# Résurrection du passé à l'aide de modèles hétérogènes d'évolution des séquences protéiques

Mathieu Groussin

N° 201-2013 Année 2013

THÈSE DE L'UNIVERSITÉ DE LYON

Présentée

devant L'UNIVERSITÉ CLAUDE BERNARD LYON 1

pour l'obtention

du DIPLÔME DE DOCTORAT
(arrêté du 7 août 2006)

soutenue publiquement le
8 novembre 2013

par

Mathieu GROUSSIN

# Résurrection du passé à l'aide de modèles hétérogènes d'évolution des séquences protéiques.

Directeur de thèse : Manolo GOUY

| | | |
|---|---|---|
| Jury : | Céline BROCHIER-ARMANET | Examinateur - Président |
| | Laurent DURET | Examinateur |
| | Nicolas GALTIER | Rapporteur |
| | Olivier GASCUEL | Rapporteur |
| | Manolo GOUY | Directeur de thèse |
| | Dominique MADERN | Examinateur |
| | Hervé PHILIPPE | Examinateur |

# UNIVERSITE CLAUDE BERNARD - LYON 1

| | |
|---|---|
| **Président de l'Université** | **M. François-Noël GILLY** |
| Vice-président du Conseil d'Administration | M. le Professeur Hamda BEN HADID |
| Vice-président du Conseil des Etudes et de la Vie Universitaire | M. le Professeur Philippe LALLE |
| Vice-président du Conseil Scientifique | M. le Professeur Germain GILLET |
| Directeur Général des Services | M. Alain HELLEU |

## *COMPOSANTES SANTE*

| | |
|---|---|
| Faculté de Médecine Lyon Est – Claude Bernard | Directeur : M. le Professeur J. ETIENNE |
| Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux | Directeur : Mme la Professeure C. BURILLON |
| Faculté d'Odontologie | Directeur : M. le Professeur D. BOURGEOIS |
| Institut des Sciences Pharmaceutiques et Biologiques | Directeur : Mme la Professeure C. VINCIGUERRA |
| Institut des Sciences et Techniques de la Réadaptation | Directeur : M. le Professeur Y. MATILLON |
| Département de formation et Centre de Recherche en Biologie Humaine | Directeur : M. le Professeur P. FARGE |

## *COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE*

| | |
|---|---|
| Faculté des Sciences et Technologies | Directeur : M. le Professeur F. DE MARCHI |
| Département Biologie | Directeur : M. le Professeur F. FLEURY |
| Département Chimie Biochimie | Directeur : Mme le Professeur H. PARROT |
| Département GEP | Directeur : M. N. SIAUVE |
| Département Informatique | Directeur : M. le Professeur S. AKKOUCHE |
| Département Mathématiques | Directeur : M. le Professeur A. GOLDMAN |
| Département Mécanique | Directeur : M. le Professeur H. BEN HADID |
| Département Physique | Directeur : Mme S. FLECK |
| Département Sciences de la Terre | Directeur : Mme la Professeure I. DANIEL |
| UFR Sciences et Techniques des Activités Physiques et Sportives | Directeur : M. C. COLLIGNON |
| Observatoire des Sciences de l'Univers de Lyon | Directeur : M. B. GUIDERDONI |
| Polytech Lyon | Directeur : M. P. FOURNIER |
| Ecole Supérieure de Chimie Physique Electronique | Directeur : M. G. PIGNAULT |
| Institut Universitaire de Technologie de Lyon 1 | Directeur : M. C. VITON |
| Institut Universitaire de Formation des Maîtres | Directeur : M. A. MOUGNIOTTE |
| Institut de Science Financière et d'Assurances | Administrateur provisoire : M. N. LEBOISNE |

**Résumé**

La reconstruction et la résurrection moléculaire de protéines ancestrales est au coeur de cette thèse. Alors que les données moléculaires fossiles sont quasi inexistantes, il est possible d'estimer quelles étaient les séquences ancestrales les plus probables le long d'un arbre phylogénétique décrivant les relations de parentés entre séquences actuelles. Avoir accès à ces séquences ancestrales permet alors de tester de nombreuses hypothèses biologiques, de la fonction des protéines ancestrales à l'adaptation des organismes à leur environnement.

Cependant, ces inférences probabilistes de séquences ancestrales sont dépendantes de modèles de substitution fournissant les probabilités de changements entre acides aminés. Ces dernières années ont vu le développement de nouveaux modèles de substitutions d'acides aminés, permettant de mieux prendre en compte les phénomènes biologiques agissant sur l'évolution des séquences protéiques. Classiquement, les modèles supposent que le processus évolutif est à la fois le même pour tous les sites d'un alignement protéique et qu'il est resté constant au cours du temps lors de l'évolution des lignées. On parle alors de modèle homogène en temps et en sites. Les modèles récents, dits hétérogènes, ont alors permis de lever ces contraintes en permettant aux sites et/ou aux lignées d'évoluer selon différents processus. Durant cette thèse, de nouveaux modèles hétérogènes en temps et sites ont été développés en Maximum de Vraisemblance. Il a notamment été montré qu'ils permettent d'améliorer considérablement l'ajustement aux données et donc de mieux prendre en compte les phénomènes régissant l'évolution des séquences protéiques afin d'estimer de meilleurs séquences ancestrales.

A l'aide de ces modèles et de reconstruction ou résurrection de protéines ancestrales en laboratoire, il a été montré que l'adaptation à la température est un déterminant majeur de la variation des taux évolutifs entre lignées d'Archées. De même, en appliquant ces modèles hétérogènes le long de l'arbre universel du vivant, il a été possible de mieux comprendre la nature du signal évolutif informant de manière non-parcimonieuse un ancêtre universel vivant à plus basse température que ses deux descendants, à savoir les ancêtres bactériens et archéens. Enfin, il a été montré que l'utilisation de tels modèles pouvait permettre d'améliorer la fonctionnalité des protéines ancestrales ressuscitées en laboratoire, ouvrant la voie à une meilleure compréhension des mécanismes évolutifs agissant sur les séquences biologiques.

**Mots-clés :** Reconstruction de séquences ancestrales, résurrection, modèles hétérogènes de substitution, température optimale de croissance, dernier ancêtre commun universel, archées, halophiles, évolution protéique.

**Abstract**

The molecular reconstruction and resurrection of ancestral proteins is the major issue tackled in this thesis manuscript. While fossil molecular data are almost nonexistent, phylogenetic methods allow to estimate what were the most likely ancestral protein sequences along a phylogenetic tree describing the relationships between extant sequences. With these ancestral sequences, several biological hypotheses can be tested, from the evolution of protein function to the inference of ancient environments in which the ancestors were adatapted.

These probabilistic estimations of ancestral sequences depend on substitution models giving the different probabilities of substitution between all pairs of amino acids. Classicaly, substitution models assume in a simplistic way that the evolutionary process remains homogeneous (constant) among sites of the multiple sequence alignment or between lineages. During the last decade, several methodological improvements were realised, with the description of substitution models allowing to account for the heterogeneity of the process among sites and in time. During my thesis, I developed new heterogeneous substitution models in Maximum Likelihood that were proved to better fit the data than any other homogeneous or heterogeneous models. I also demonstrated their better performance regarding the accuracy of ancestral sequence reconstruction.

With the use of these models to reconstruct or resurrect ancestral proteins, my coworkers and I showed the adapation to temperature is a major determinant of evolutionary rates in Archaea. Furthermore, we also deciphed the nature of the phylogenetic signal informing substitution models to infer a non-parsimonious scenario for the adaptation to temperature during early Life on Earth, with a non-hyperthermophilic last universal common ancestor living at lower temperatures than its two descendants. Finally, we showed that the use of heterogeneous models allow to improve the functionality of resurrected proteins, opening the way to a better understanding of evolutionary mechanisms acting on biological sequences.

**Keywords:** Ancestral sequence reconstruction, resurrection, heterogeneous substitution models, otpimal growth temperature, last universal common ancestor, archaea, halophiles, protein evolution.

# Remerciements

Ma mémoire est défaillante. Cela doit être l'heure. Ou autre chose. Je ne me souviens plus très bien à partir de quand cela a commencé. Mais il me semble me souvenir que j'ai commencé il y a environ 15 ans. Après avoir piqué dans le bureau de mes deux parents des stylos et des copies, je m'amusais à jouer au prof. Comme mes parents. Qui n'avaient eux, pas tout le temps l'air de s'amuser. Mais j'adorais ça. Pendant longtemps j'ai voulu être prof, faire ce métier formidable à bien des égards. L'insouciance et la naïveté de la jeunesse me direz vous. Quoiqu'il en soit, cela a influencé tous mes choix d'orientation durant mon parcours scolaire, jusqu'à rentrer à l'ENS de Lyon dans l'unique but de pouvoir intégrer sa préparation à l'agrégation. C'était sans compter ma rencontre avec trois personnes, trois enseignants-chercheurs, qui ont quand même, il faut bien le dire, foutu un sacré bordel. Je me suis tellement passionné pour ce qu'ils racontaient que j'ai laissé tomber toutes mes plus profondes envies concernant mon futur métier. Je remercie ici Marie Sémon, Ludovic Orlando et Manolo Gouy pour avoir rendu une partie des enseignements de Licence terriblement passionnants.

Manolo, merci pour tout. Il n'y a pas grand chose à dire. Ou au contraire il y en aurait beaucoup trop. J'ai tant appris à tes côtés durant ces quatres ans que j'espère du fond du coeur pouvoir continuer à le faire dans l'avenir. Tu fais partie de ces gens capables de valoriser en permanence la personne avec qui tu intéragit scientifiquement, tout en conservant la rigueur nécessaire à la construction du raisonnement. C'est inestimable. Tant du point de vue professionnel que personnel. Bon malgré tout, tu n'as pas tout réussi : je crois que je continuerai à écrire et dire *résurrecter*. J'y arrive pas, j'y arrive pas !

Je remercie Dominique Mouchiroud, pour ses cours extrêmement passionnants sur l'évolution moléculaire dispensés pendant ma première année de Master, ainsi que pour m'avoir permis d'effectuer ma thèse dans son laboratoire. Dominique, ce fut un honneur de te cotoyer au LBBE. Et des personnes avec lesquelles c'est un *honneur* de travailler, il n'y en a pas forcément *légion*.

Je remercie mes compagnons de routes phylogénétiques, au côté de qui travailler n'est pas une contrainte, mais un plaisir constant. Merci donc à Vincent, Éric, Gergely et Bastien. Merci de m'avoir familiarisé avec le concept de corrélation positive entre le côté inutile et le côté passionnant d'une conversation. On a souvent refait le monde et trouvé un plaisir extrêment cynique et délicieux à voir certaines personnes se débattre avec leur destin malheureux. Vivement qu'on s'y remette. Ah au fait, Vincent : promis en post-doc je prendrai une demi-heure par jour pour te faire découvrir la musique 2.0.

Je remercie chaleureusement l'ensemble des membres de mon jury de thèse, pour avoir accepter d'évaluer à la fois mon manuscrit écrit et ma soutenance orale. Ce fut un honneur d'être évaluer par Céline Brochier-Armanet, Laurent Duret, Nicolas Galtier, Olivier Gascuel,

Dominique Madern et Hervé Philippe. Je vous suis extrêmement reconnaissant pour vos conseils et vos remarques, qui ont permis de grandement améliorer le manuscrit.

Merci à tous ceux avec qui j'ai collaboré directement durant cette thèse et avec qui j'ai pris un immense plaisir à apprendre : Bastien et Gergely de nouveau, Céline pour m'avoir proposé de participer à un de ses projets, Laurent G., Nicolas L., Samuel, Sandrine, Joanne, Vic, Dominique, Ziheng et enfin Julien D. Un merci tout particulier à mes compagnons de debuggages, Thomas, Laurent G. et Julien D., pour toute leur aide très précieuse. Merci aussi à tous ceux qui, au travers d'échanges divers, scientifiques ou lubriques, ont fait en sorte de contribuer à mon bien être au LBBE. Pour le côté scientifique : Yann, Thomas, Céline, Laurent D., Sylvain, Floriane, Nicolas R., Eugénie, Raquel, Gab, Dominique G. Marc, Jean, Anne-Béatrice, Daniel, Florent, Rémi, Matthieu, Erika, Franck, Bérénice, Vincent M., Nicolas R., Philippe. Pour le côté lubrique : Yann, Thomas, Florent, Rémi, Sylvain, Magali, Floriane, Joanna, Héloïse, Aline, Marie, Fanny, Bérénice, Dominique G., Philippe. Il y en a quand même pas mal qui sont des deux côtés...

Merci aux quatres mecs ingénieux d'en face. Je les ai souvent embêtés, ils étaient toujours de bonne humeur. J'ai souvent rencontré des problèmes, ils avaient toujours des solutions. Merci, d'abord par galanterie puis par ordre alphabétique, à Simon, Bruno, Lionel et Stéphane. Bien qu'il trouve que le *revival* actuel de la Cold et de la New Wave ne soit pas capable de produire des groupes à la hauteur de ceux qu'il a pu voir en concerts à la fin des 1970, début des années 1980, discuter musique avec Guy est extrêmement passionnant. Merci à lui pour les découvertes qu'il m'a permis de faire.

Merci aux quatres filles d'en bas. Leur aide est toujours précieuse, et leur constante disponibilité est à la hauteur de leur bonne humeur. Big up à Aline, LaetiCia, Nathalie et Odile.

Je remercie l'ensemble de ma famille. Je n'ai pas l'occasion de vous voir souvent, mais je pense souvent à vous. Je remercie tout particulièrement mes parents, pour leur soutien indéfectible tout au long de mon parcours et leur éducation centrée sur une certaine idée de l'humain. Papa, mon introduction aborde partiellement des concepts religieux. Même si je sais que cela va être dur pour toi après t'avoir offert *Pour en finir avec Dieu* de Dawkins, ne m'en veux pas. Et s'il te plait, ne me déshérite pas. Tu sais, c'est une façon comme une autre d'essayer d'être transgressif, de ne pas suivre l'ordre établi qui souhaiterait que j'introduise ma thèse portant sur l'évolution en présentant le vieux barbu anglais et sa vie soit-disante passionnante. Et s'il y a, pour moi, un modèle de tête de mule transgressive, c'est bien toi. Alors merci. Maman, merci pour m'avoir montré un soir sur internet ce qu'était l'Ecole Normale Supérieure. Il est fort probable que je n'aurais pas eu ce parcours si ce soir là, au lieu de venir discuter de mon avenir avec moi, tu avais choisi de vaquer à tes propres occupations. Cher frère, merci aussi pour ton soutien. Je suis fier de toi et de ton parcours. Continue, j'aurai besoin de toi plus tard si je continue de courir autant.

Pour les meilleurs concerts, pour les meilleures nuits blanches, pour les meilleures aventures, pour les meilleurs mojitos, pour les meilleurs cookies, pour les meilleures bouffes gargantuesques, pour les meilleurs soutiens, merci à Chloé, Pierre L., Yann, Pierre L.M., Blaise, Florent, Florie. N'interprétez pas l'ordre d'apparition de votre nom dans cette liste...

Il y a quatre mecs que je tiens à remercier tout spécialement. Trois bretons et un Villeurbannais. Ils se reconnnaitront. Pour tout ce que l'on a échangé et vécu ensemble, merci. Je traverserais le monde pour vous. Enfin, faites pas les idiots hein, attendez au moins que j'accumule quelques miles en faisant des aller-retours au dessus de l'Atlantique dans les mois qui viennent.

*Cette thèse est dédiée à Dietrich Mateschitz et Chaleo Yoovidhya, sans qui je n'aurais jamais pu accomplir la rédaction à temps.*

*Notre tête est ronde pour permettre à la pensée de changer de direction.*
**Francis Picabia**

*Sans la musique, la vie serait une erreur.*
**Friedrich Nietzsche**

# Table des matières

# 1

# Introduction.

## 1.1 Introduction aux différents concepts de Résurrection.

La *résurrection* est définie comme le fait de revenir à la vie après la mort. C'est un concept qui peut très bien s'appliquer à un organisme vivant, une idée, un système politique etc. Ce concept a inondé de nombreuses croyances et religions par le passé et est toujours actuellement l'objet de fascinations pouvant prendre des formes diverses et variées. Une distinction est faite entre la résurrection et la renaissance. La résurrection suppose que l'individu, l'idée etc reprend vie dans une configuration de lieu et physique identiques à celles qu'il ou elle possédait avant la mort. La renaissance quant à elle conçoit également un retour à la vie après la mort mais dans un cadre différent de celui originel. Cette renaissance peut s'effectuer sous une forme différente où avoir lieu dans un autre monde. Alors que dans les religions bouddhiste ou hindouiste la renaissance peut être associée à une punition de l'âme ou du corps suite à des péchés physiques ou moraux, les religions égyptienne antique, chrétienne et juive concoivent généralement la résurrection comme quelque chose de positif. Elle est systématiquement associée au retour à la vie de l'être, par le passage physique du corps de la position allongée (suite à sa mort) à la position levée. Chez les égyptiens, le mythe d'Osiris est souvent associé à une résurrection. Après être devenu

successivement le Dieu de la Fertilité et du développement végétal en passant par le Dieu des Morts, Osiris finit par être le Dieu de la Résurrection à la Vie Eternelle après avoir été tué par le Dieu Seth, son frère jaloux. Le culte d'Osiris et sa mythologie se sont développés jusqu'à occuper le champ majeur de la religion funéraire. Le concept de résurrection associé à Osiris est malgré tout assez controversé. En effet, il est souvent mentionné qu'Osiris a été le sujet d'une renaissance car il est ranimé dans le royaume de l'Au-delà dont il devient le souverain. Une des images de la résurrection d'Osiris est la figuration d'épis de céréales poussant sur son corps momifié. Le cycle annuel de la végétation qui meurt puis renaît symbolise le concept de résurrection.

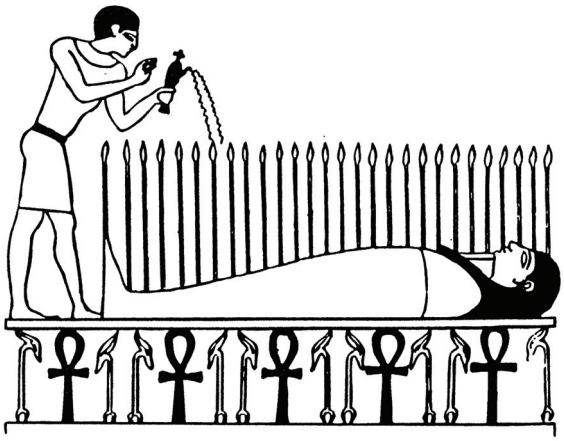


FIGURE 1.1 – Epis de céréales poussant sur le corps momifié d'Osiris, symbole de résurrection.

Chez le Christianisme, la résurrection est mentionnée dans de nombreux écrits comme l'Ancien Testament, les Évangiles ou les Actes de Apôtres. La plus importante concerne bien évidemment Jésus, dont la résurrection n'est pas un simple retour à la vie physique sur Terre mais le passage à la vie nouvelle en Dieu. D'après les Évangiles, Jésus a ressuscité trois jours après sa mort, la matin de Pâques. La religion musulmane place également la résurrection dans une position d'importance. Il est considéré qu'Allah ramènera tous les hommes à la vie après leur mort le Jour Promis. Les Chiites, représentant une branche de la religion musulmane, croient quant à eux que la résurrection se manifestera par le retour des morts sous leur forme et avec leur corps et leur âme originels.

Au delà du cercle des religions, la notion de résurrection est présente dans de nombreuses manifestations artistiques. De nombreux romans, bandes dessinées, films ou épisodes de séries télévisées portent le nom de *Résurrection*. De même, de nombreux artistes musicaux ont produit des albums dont le titre est *Résurrection* ou *Resurrection* en anglais. Par exemple, *Resurrection* est un album à titre posthume du maitre absolu du Hip-Hop, 2pac, sorti en 2003. Virgo Four, duo

anglais de musique house-techno a également sorti en 2011 un album portant ce nom. Enfin, Ophélie Winter a, en 2009, sorti un album portant ce nom, symbolisant son retour tant attendu à la chanson, rompant le désespoir de nombreux francophones ayant perdu la foi, depuis 2002 et son précédent album, de pouvoir revoir un jour Ophélie sur le devant de la scène.

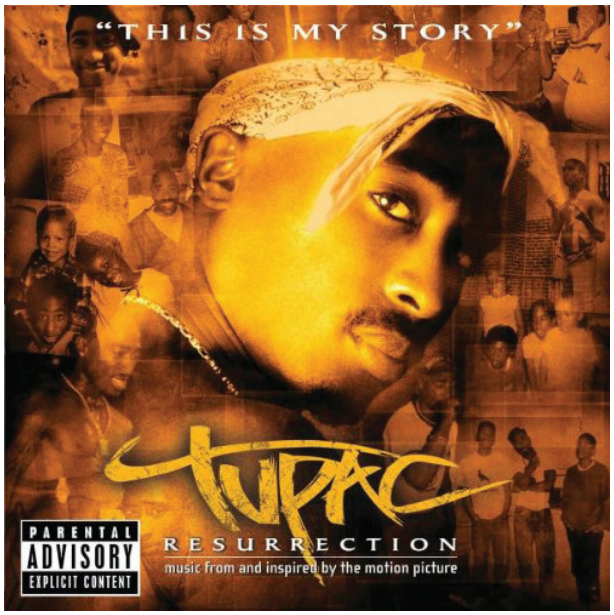



FIGURE 1.2 – *Resurrection*. Album posthume de 2pac.

La thèse que je présente ici place la renaissance et la résurrection au centre du sujet. Cette résurrection sera abordée d'un point de vue scientifique, à l'aide d'une démarche rationnelle d'observation et d'interprétation des expériences, conférant à la démarche scientifique sa nécessaire propriété de répétabilité des expériences. La suite de mon message se place entièrement dans ce cadre rigoureux. Dans le film Jurassic Park, des dinosaures sont ressuscités à la suite de clonages de leur propre ADN, découvert intact à l'intérieur d'intestins de moustiques préservés dans de l'ambre. La question est de savoir s'il est possible un jour d'assister à un tel scénario où, à partir d'ADN ancien et entièrement préservé, une espèce éteinte est ramenée à la vie. Bien que la science du clonage ne soit encore qu'à ses prémices, il n'apparait pas *a priori* inenvisageable de réaliser une telle prouesse méthodologique, dans un avenir plus ou moins proche. Ressusciter une espèce nécessiterait néanmoins plusieurs résurrections d'individus, de sexes opposés dans le cas d'une espèce sexuée, et capables de se reproduire. Au delà des questions éthiques et philosophiques qu'une telle démarche scientifique implique, il est raisonnable de se poser la question de l'intérêt de ressusciter un organisme ou une espèce éteinte. Cette question ne sera pas abordée ici.

Une des difficultés majeures est de pouvoir avoir accès aux génomes conservés d'espèces éteintes, ce qui limite grandement la liste des candidats potentiels à la résurrection (l'ADN se dégrade rapidement au cours du temps, dû aux rayons UV, au processus de déamination etc). Malgré tout, il existe quelques espèces éteintes mythiques dans l'esprit des gens et auxquelles une attention peut être portée. Par exemple, le Mammouth (*Mammuthus primigenius*) représente une cible intéressante : l'espèce s'est éteinte récemment et plusieurs spéciments fossiles entiers ont été retrouvés dans la toundra arctique. Une collaboration internationale constituée de scientifiques russes et coréens se penche actuellement sur la question de la résurrection du Mammouth, dont l'ADN ancien pourrait être cloné à l'aide d'une éléphante (Zimmer, 2013). En outre, des espèces très récemment éteintes dont des spécimens ont été conservés dans des musées représentent également des candidats à la résurrection. Des prélèvements peuvent être réalisés sur les tissus anciens, afin d'en extraire l'ADN et de tenter de le cloner dans une espèce proche. Le tigre de Tasmanie, endémique à l'Australie, est un animal fascinant car il fut un des plus gros marsupiaux carnivores ayant fréquenté l'homme. L'espèce s'est éteinte dans les années 1930 par la volonté sans relâche des chasseurs de le voir disparaître. Malgré cela, des spécimens furent conservés dans des musées australiens. Enfin, le dodo est une des espèces éteintes les plus célèbres dans l'opinion publique. Cet oiseau vivait sur l'île Maurice et plusieurs spécimens sont également conservés dans des musées.

FIGURE 1.3 – Espèces éteintes pour lesquelles une résurrection peut être envisageable. A gauche, le Mammouth (*Mammuthus primigenius*) ; au centre, le tigre de Tasmanie (*Thylacinus cynocephalus*) ; à droite, le dodo (*Raphus cucullatus*)

Une espèce a été le sujet des premières tentatives de résurrection. Il s'agit du bouquetin d'Espagne ou bouquetin ibérique (*Capra pyrenaica*). Quatre sous-espèces exitent chez cette espèces. Une d'entre elle s'est éteinte durant le XIX$^{eme}$ siècle, tandis qu'une autre s'est éteinte à la fin des années 1990. Alors que cette sous-espèce (*Capra pyrenaica ssp. pyrenaica*) était en danger, des scientifiques ont entrepris des prélèvements d'ADN sur des individus encore vivants afin d'envisager un clonage dans le futur. Les différentes tentatives de clonage à l'aide d'ovocytes

de chèvre ont toutes échoué, entrainant l'arrêt du projet (Pina-Aguilar et al., 2009).

Récemment, plusieurs génomes complets anciens ont pu être séquencés. Parmi ceux là, les génomes de Néandertal (Green et al., 2010) et de Denisova (Reich et al., 2010) ont permis d'avoir accès à des séquences génomiques ancestrales le long de la branche humaine, ouvrant la voie à une meilleure compréhension de la dynamique évolutive de notre génome. A une autre échelle, le génome de souches de bactéries *Yersinia pestis* responsables de l'épidémie de peste noire ayant eu lieu au XIV$^{eme}$ siècle a été séquencé, suite à des prélèvements sur des os et des dents de victimes de l'épidémie (Bos et al., 2011), en espérant mieux comprendre les dynamiques évolutives entrainant l'apparition d'agents infectieux hautement pathogènes. Ces dernières prouesses méthodologiques permettant de faire renaître des génomes anciens illustrent les capacités des biotechnologies à offrir un matériel d'étude inespéré à la communauté scientifique. Malgré tout, la question de ressusciter Néandertal, Denisova ou la peste noire à partir de ces génomes n'a pas été encore envisagée sérieusement. Heureusement ?

La présence dans le registre fossile de restes d'organismes permettant d'avoir accès à des séquences moléculaires ancestrales (ADN ou protéines) est très limitée. Si cela était possible, cela permettrait de mieux comprendre comment les séquences moléculaires, et notamment les protéines, évoluent au cours du temps et acquièrent de nouvelles fonctions/structures. Il serait même éventuellement possible de relier ces événements évolutifs à des changements adaptatifs des espèces ancestrales portant ces molécules. Mais il est possible de contourner ces limitations, en exploitant la richesse d'information présente dans les séquences protéiques actuelles. Pour celles qui ont conservé le signal évolutif informant le chemin substitutionnel emprunté par leurs ancêtres, l'utilisation de modèles probabilistes d'évolution des séquences moléculaires peut alors permettre, dans un premier temps par ordinateur, puis éventuellement en laboratoire, d'estimer et de synthétiser les séquences protéiques ancestrales aux séquences actuelles. La reconstruction (ou renaissance) *in silico*, évenuellement suivie d'une résurrection moléculaire *in vitro* ou *in vivo* a, dès 1963, été envisagée par Pauling and Zuckerkandl. C'est maintenant devenu un champ de recherche à part entière dans lequel cette thèse s'inscrit modestement.

## 1.2 Principes généraux de la modélisation de l'évolution moléculaire.

Au cours du temps et au sein d'une espèce ou d'une population, les séquences d'ADN et de protéines évoluent d'une génération à l'autre, en accumulant des mutations, dont le destin (élimination, fixation, maintien à une fréquence intermédiaire) peut être déterminé par des processus sélectifs ou neutres. Les questions que les sections 1.2 et 1.3 abordent concernent l'explication

de la façon dont l'évolution des séquences est modélisée mathématiquement et comment, à partir de cette modélisation, le calcul d'un arbre phylogénétique retraçant l'histoire évolutive des séquences est réalisé.

### 1.2.1 Modèles de Markov

**Remarque :** Cette section, concernant les modèles markoviens, leur utilisation en phylogénie et la description des modèles nucléiques et de codons, se veut volontairement succinte. L'accent sera plus mis sur les modèles protéiques, qui sont au coeur de cette thèse. Cependant, la présentation des principes fondamentaux nécessaires à la compréhension et l'utilisation de ces modèles est effectuée.

#### 1.2.1.1 Définition mathématique

Un processus markovien d'ordre $r$ est un processus stochastique qui possède la propriété de Markov, stipulant que la distribution conditionnelle des états futurs ne dépend que des $r$ distributions des états présents et passés. Un processus de Markov d'ordre 1 stipule ainsi que la distribution conditionnelle des états futurs ne dépend que de la distribution des états présents et non de ceux du passé. C'est alors un processus sans mémoire. Il se définit mathématiquement de la façon suivante :
soit $(X_{n+1}) = X_1, X_2, \ldots, X_n, X_{n+1}$ une suite de v.a. dont les états possibles appartiennent à un ensemble $E$ fini ou dénombrable. Cette suite constitue une *chaine de Markov d'ordre r* si :

$$\begin{aligned}&\mathbb{P}(X_{n+1} = i_{n+1} | X_n = i_n, \ldots, X_2 = i_2, X_1 = i_1)\\ &= \mathbb{P}(X_{n+1} = i_{n+1} | X_n = i_n, \ldots, X_{n-r+1} = i_{n-r+1})\end{aligned}$$

Dans le cas d'une *chaine de Markov d'ordre 1*, on a :

$$\begin{aligned}&\mathbb{P}(X_{n+1} = i_{n+1} | X_n = i_n, \ldots, X_2 = i_2, X_1 = i_1)\\ &= \mathbb{P}(X_{n+1} = i_{n+1} | X_n = i_n)\end{aligned}$$

La modélisation des changements de bases (ou d'acides aminés) au cours du temps va alors être permise à l'aide d'un processus markovien d'ordre 1. En effet, les séquences évoluent par l'effet de mécanismes de mutation et sélection, qui n'ont pas accès aux états passés. Les différentes probabilités du processus décrivant les passages d'un état de la chaine à un autre modélisent alors les mécanismes biologiques entrainant l'évolution des séquences.

#### 1.2.1.2 Homogénéité, stationnarité et réversibilité

Une chaîne de Markov d'ordre 1 est dite *homogène en temps* si :

$$\forall i, j \in E, \quad \exists \quad p_{ij}, \quad \forall n \in \mathbb{N} \quad \text{tel que} \quad \mathbb{P}(X_{n+1} = j | X_n = i) = p_{ij}$$

**Dans le cas homogène, la probabilité de passer de $i \rightarrow j$ ne dépend que de $i$ et $j$ quelle que soit l'étape considérée.** Le processus est alors considéré comme constant au cours du temps, quelle que soit l'échelle évolutive des séquences/espèces considérées.

- Les valeurs $p_{ij} = \mathbb{P}(X_{n+1} = j | X_n = i)$ constituent les *probabilités de transition* de la chaine de Markov entre les états possibles.
- On appelle *matrice de transition*, la matrice $P = (p_{ij})$ contenant l'ensemble des valeurs de $p_{ij}$ possibles :

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1s} \\ p_{21} & p_{22} & \cdots & p_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ p_{s1} & p_{s2} & \cdots & p_{ss} \end{pmatrix} \quad \begin{array}{c} E = \{1, 2, \ldots, s\} \\ 0 \leq p_{ij} \leq 1 \\ \sum_j p_{ij} = 1 \end{array}$$

- La matrice $P^{(n)}$ des probabilités pour $n$ étapes de la chaine de Markov est égale à $P^n$.
- L'état du système après ces $n$ transitions est donnée par le vecteur de probabilités $F^{(n)}$ :

$$F^{(n)} = \left(f_1^{(n)}, f_2^{(n)}, \ldots, f_s^{(n)}\right)$$

tel que :

$$F^{(n)} = F^{(n-1)}P = F^{(0)}P^n$$

avec $F^{(0)}$ la distribution initiale.

- Lorsque $n \rightarrow \infty$, $F^{(n)}$ converge vers une distribution invariante (également appelée *stationnaire*) notée $\Pi = (\pi_i)$, telle que :

$$\Pi P = \Pi$$

Alors, chaque ligne de la matrice $P^n \rightarrow \Pi$.

- Toute chaîne de Markov possède au moins une distribution invariante. Dans la suite du manuscrit, cette distribution stationnaire sera appelée **profil**.

Dans le cas d'une séquence dont l'évolution moléculaire est modélisée par un processus markovien, **le profil du processus représente ainsi la composition globale (en bases, codons ou acides aminés) à l'échelle de toute la séquence lorsque l'équilibre est atteint**. Si la séquence n'est pas à l'équilibre compositionnel (voir plus bas), le profil spécifie ainsi la direction vers laquelle la composition de la séquence évolue.

- Un processus de Markov stationnaire est dit *réversible* dans le temps si

$$\pi_i p_{ij} = \pi_j p_{ji}$$

  pour chaque paire d'états $(i, j)$. Avec $\pi_i$ et $\pi_j$, les fréquences stationnaires des caractères $i$ et $j$ dans la séquence, et $p_{ij}$ et $p_{ji}$, les probabilités de transition de $i \to j$ et de $j \to i$.

- **La réversibilité signifie que lorsque l'équilibre est atteint, la probabilité ou quantité espérée de changements de l'état $i$ vers $j$ est égale à la probabilité ou quantité espérée de changements de l'état $j$ vers l'état $i$**

### 1.2.2 D'un temps discret au temps continu

Jusque ici, la chaîne de Markov considérait des variables aléatoires en temps discret. Cependant, les molécules évoluent dans un temps qui est continu. Il est donc nécessaire de généraliser le fonctionnement du processus markovien au temps continu. Dans ce cas, on a :

$$\text{Lorsque } h \to 0 : \quad \mathbb{P}(X(t+h) = j | X(t) = i) \simeq q_{ij} h$$

Les valeurs de $q_{ij}$ définissent la matrice $Q = (q_{ij})$, dite des *taux instantanés*, telle que :

$$Q = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1s} \\ q_{21} & q_{22} & \cdots & q_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ q_{s1} & q_{s2} & \cdots & q_{ss} \end{pmatrix} \quad \begin{array}{c} q_{ij} \geq 0 \ (i \neq j) \\ q_{ii} = -\sum_{j \neq i} q_{ij} \\ \sum_j q_{ij} = 0 \end{array}$$

Dans le cas d'un modèle de Markov appliqué aux séquences d'ADN, on a alors :

$$Q = \begin{pmatrix} \mu_{AA} & \mu_{AT} & \mu_{AC} & \mu_{AG} \\ \mu_{TA} & \mu_{TT} & \mu_{TC} & \mu_{TG} \\ \mu_{CA} & \mu_{CT} & \mu_{CC} & \mu_{CG} \\ \mu_{GA} & \mu_{GT} & \mu_{GC} & \mu_{GG} \end{pmatrix} = \begin{pmatrix} -\lambda_A & \mu_{AT} & \mu_{AC} & \mu_{AG} \\ \mu_{TA} & -\lambda_T & \mu_{TC} & \mu_{TG} \\ \mu_{CA} & \mu_{CT} & -\lambda_C & \mu_{CG} \\ \mu_{GA} & \mu_{GT} & \mu_{GC} & -\lambda_G \end{pmatrix}$$

avec :

- $\mu_{ij}$ $(i \neq j)$ le *taux de substitution instantané* d'un nucléotide de l'état $i$ vers l'état $j$,
- $\lambda_i$ le *taux de changement instantané* d'un nucléotide dans l'état $i$ vers un autre nucléotide
- et, par conséquent, $1 - \lambda_i$ le *taux de conservation instantané* d'un nucléotide.

Toutes ces notations sont généralisables à un alphabet de codons ou d'acides aminés en considérant un ensemble $E$ de dimension 61 ou 20, respectivement.

À présent, la figure 1.4 présente les scénarios possibles pour atteindre l'état A au temps $t+dt$ et va nous permettre de modéliser de manière générale la dynamique des fréquences des états au cours du temps entre $t$ et $t + dt$ :

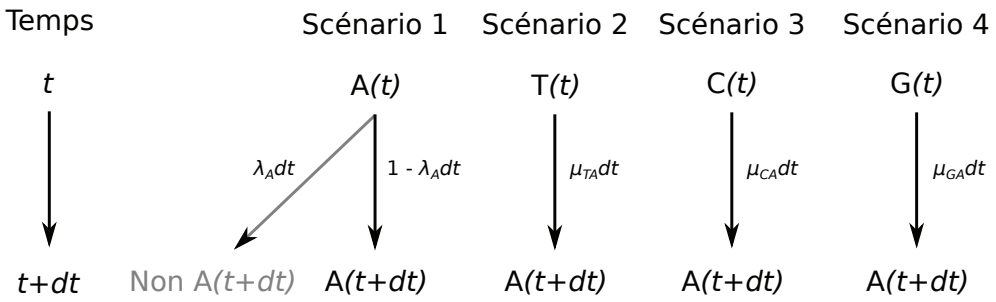



FIGURE 1.4 – Intégration des chemins possibles pour arriver à l'état A au temps $t + dt$. Cette figure est inspirée de la figure 2.2 du livre de Perrière and Brochier-Armanet (2010).

À partir de cette figure, il est aisé de voir apparaitre le système de quatre équations différentielles suivant :

$$\begin{cases} A(t+dt) = A(t)(1 - \lambda_A dt) + T(t)\mu_{TA}dt + C(t)\mu_{CA}dt + G(t)\mu_{GA}dt \\ T(t+dt) = T(t)(1 - \lambda_T dt) + A(t)\mu_{AT}dt + C(t)\mu_{CT}dt + G(t)\mu_{GT}dt \\ C(t+dt) = C(t)(1 - \lambda_C dt) + A(t)\mu_{AC}dt + T(t)\mu_{TC}dt + G(t)\mu_{GC}dt \\ G(t+dt) = G(t)(1 - \lambda_G dt) + A(t)\mu_{AG}dt + T(t)\mu_{TG}dt + C(t)\mu_{CG}dt \end{cases}$$

Soit $F(t)$ le vecteur des fréquences des quatre bases au temps $t$ tel que $F(t) = \{A(t), T(t), C(t), G(t)\}$. En écriture matricielle, le système des quatre équations différentielles précédentes devient donc :

$$F(t + dt) = F(t) + {}^{t}QF(t)dt$$

soit :

$$\frac{dF(t)}{dt} = {}^{t}QF(t) \tag{1.1}$$

qui est l'équation générale du processus de Markov considéré en temps continu, dont la solution est :

$$F(t) = F(0)e^{{}^{t}Qt}$$

avec $F(0) = \{A(0), T(0), C(0), G(0)\}$ le vecteur des fréquences ancestrales.

On définit alors la matrice $P(t)$, matrice des probabilités de transitions entre états continus, en fonction de la matrice des taux instantanés :

$$P(t) = e^{Qt} = \begin{pmatrix} p_{AA}(t) & p_{AC}(t) & p_{AT}(t) & p_{AG}(t) \\ p_{CA}(t) & p_{CC}(t) & p_{CT}(t) & p_{CG}(t) \\ p_{TA}(t) & p_{TC}(t) & p_{TT}(t) & p_{TG}(t) \\ p_{GA}(t) & p_{GC}(t) & p_{GT}(t) & p_{GG}(t) \end{pmatrix}$$

Le calcul de la matrice $P(t)$ nécessite le calcul d'une exponentielle de matrice, ce qui peut se réaliser assez aisément en calculant les valeurs et vecteurs propres la matrice $Q$ dans le cas où celle-ci est diagonalisable. Si ce n'est pas réalisable analytiquement, la matrice $P(t)$ peut être assez bien approchée par un développement de Taylor à faible degré. L'explication de ces calculs dépasse le cadre de cette présentation brève des modèles markoviens appliqués à l'évolution moléculaire.

### 1.2.3 Modèles nucléiques

Plusieurs modèles nucléiques ont été proposés depuis la fin des années 70 et le modèle de Jukes et Cantor (Jukes and Cantor, 1969). Les différents modèles visent à prendre en compte les spécificités de l'évolution moléculaire des bases, comme la différence entre les taux de transition et les taux de transversions, les fréquences d'équilibre différentes entre bases. En fonction de la complexité des modèles, les probabilités de transitions entre états peuvent ou non être exprimées analytiquement. La présentation précise de tous ces modèles dépasse le cadre de cette thèse. Les

livres de Felsenstein (2004) et Yang (2006) expliquent très clairement l'ensemble de ces modèles et de leurs implications. Néanmoins, il est possible de citer les modèles suivants :

- Jukes & Cantor à 1 paramètre (JC69).
- Kimura à deux paramètres (K81).
- Felsenstein à trois paramètres (F81).
- Hasegawa, Kishino & Yano à quatre paramètres (HKY85).
- Felsenstein 1984 à cinq paramètres (F84).
- Tamura à trois paramètres (T92).
- Tamura & Nei à six paramètres (TN93).
- *Generalised Time Reversible* à neuf paramètres (GTR).

Dans le comptage du nombre de paramètres pour chacun de ces modèles, le paramètre de taux $r$ est considéré. Ce paramètre peut s'éliminer dans le cas où l'on impose une normalisation de la matrice afin de contraindre le taux moyen global à une valeur donnée (par exemple un taux global de 1 afin de rendre interprétable les longueurs de branches d'un arbre phylogénétique en nombre de substitution moyen par site – voir plus loin, section Modèles protéiques). Le modèle GTR est le modèle réversible le plus général et peut être étendu à un modèle de codons ou d'acides aminés. Il considère que tous les taux de substitutions instantanés entre états sont différents et peuvent être estimés en fonction des données, par exemple par Maximum de Vraisemblance (voir section 1.3). Le modèle GTR est riche en paramètres, puisqu'il possède par défaut 9 paramètres dans le cas de l'ADN. En effet, d'après l'hypothèse de réversibilité ($\pi_i \mu_{ij} = \pi_j \mu_{ji}$), on a :

$$\mu_{ij} = \frac{\mu_{ji}}{\pi_i} \pi_j = \rho_{ij} \pi_j$$

avec $\rho_{ij}$ nommé *échangeabilité* de $i$ vers $j$. L'échangeabilité $\rho_{ij}$ représente la propension de l'état $i$ à évoluer vers l'état $j$. On a alors $\rho_{ij} = \rho_{ji}$, d'où :

$$Q = \begin{pmatrix} \mu_{AA} & \mu_{AC} & \mu_{AT} & \mu_{AG} \\ \mu_{CA} & \mu_{CC} & \mu_{CT} & \mu_{CG} \\ \mu_{TA} & \mu_{TC} & \mu_{TT} & \mu_{TG} \\ \mu_{GA} & \mu_{GC} & \mu_{GT} & \mu_{GG} \end{pmatrix} = \begin{pmatrix} -\lambda_A & \rho_{AC}\pi_C & \rho_{AT}\pi_T & \rho_{AG}\pi_G \\ \rho_{AC}\pi_A & -\lambda_C & \rho_{CT}\pi_T & \rho_{CG}\pi_G \\ \rho_{AT}\pi_A & \rho_{CT}\pi_C & -\lambda_T & \rho_{TG}\pi_G \\ \rho_{AG}\pi_A & \rho_{CG}\pi_C & \rho_{TG}\pi_T & -\lambda_G \end{pmatrix}$$

Soit :

$$Q = S \times diag(\Pi)$$

avec :

- $S$ la matrice symétrique des échangeabilités,
- $\rho_{ii} = \frac{-\lambda_i}{\pi_i}$,
- et $diag(\Pi)$ la matrice diagonale contenant les fréquences stationnaires, ou d'équilibre des états de la chaine de Markov.

La matrice $\Pi$ est souvent mentionnée comme un vecteur de fréquences d'équilibre, ou profil. Dans le cas nucléique, il y a alors 6 échangeabilités à estimer, et trois fréquences d'équilibres, soient 9 paramètres.

### 1.2.4 Modèles de codons

L'évolution des séquences nucléotidiques codantes peut aussi être modélisée à l'aide de modèles markoviens de substitution de codons. Dans ce cas, les différents états de la chaine de Markov ne sont plus les quatre nucléotides mais les triplets de nucléotides codant pour un acide aminé, ou codon. Il existe en tout 64 codons, dont trois codons stop, pour 20 acides aminés (dans le cas général). Par conséquent, plusieurs codons peuvent coder le même acide aminé, ce qui donne au code génétique sa propriété dite de redondance. Une conséquence de cette propriété de redondance est qu'il devient possible de distinguer les substitutions nucléotidiques synonymes, qui n'entrainent pas de changement d'acide aminé des substitutions non-synonymes, qui provoquent un changement d'acide aminé. Étant donné que les forces sélectives agissent principalement au niveau protéique, les pressions de sélection sont radicalement différentes selon que la mutation est synonyme ou non-synonyme. La comparaison des taux de substitution synonymes vis à vis des taux de substitutions non-synonymes donne ainsi accès à une mesure efficace des effets de la sélection, et peut permettre de détecter des lignées ou des sites particuliers qui sont sous des régimes sélectifs non-neutres, c'est à dire sur lesquels la sélection naturelle opère (Nielsen and Yang, 1998; Yang, 2006).

Les premiers modèles de codons furent proposés par Goldman and Yang (1994) et Muse and Gaut (1994). La chaine de Markov est utilisée pour modéliser le processus de substitutions entre codons. Dans ces modèles, en considérant le paramètre $\omega$ comme étant le rapport des taux de substitutions non-synonymes sur synonymes (communément appelé dN/dS), $\kappa$ le rapport

des taux de transitions sur transversions et $\pi_j$ la fréquence d'équilibre du codon $j$, le taux de substitution instantané entre codons $i$ et $j$ est défini comme suit :

$$q_{ij} = \begin{cases} 0, & \text{si } i \text{ and } j \text{ diffèrent à deux ou trois positions de codon,} \\ \pi_j, & \text{si } i \text{ and } j \text{ diffèrent par une transversion synonyme,} \\ \kappa\pi_j, & \text{si } i \text{ and } j \text{ diffèrent par une transition synonyme,} \\ \omega\pi_j, & \text{si } i \text{ and } j \text{ diffèrent par une transversion non-synonyme,} \\ \omega\kappa\pi_j, & \text{si } i \text{ and } j \text{ diffèrent par une transition non-synonyme.} \end{cases}$$

Des modèles de codons plus sophisitiqués ont par la suite été développés, sur la base de modèles de mélange de sous-modèles. Ces sous-modèles peuvent varier selon la valeur d'$\omega$ qui peut être soit fixé à 0 (sélection purificatrice) ou 1 (sélection neutre), soit estimé à partir des données afin de détecter des sites ou des lignées sous pression de sélection positive (Nielsen and Yang, 1998; Yang et al., 2000; Yang and Nielsen, 2002).

### 1.2.5 Modèles protéiques

Avant l'avènement des techniques de séquençage permettant d'avoir accès à une grande quantité de données nucléiques (pouvant ensuite être traduite en données protéiques pour les parties codantes), les premières analyses phylogénétiques ont utilisé des séquences protéiques. Ainsi, le premier arbre phylogénétique moléculaire reconstruit était basé sur l'étude de sept séquences de 19 acides aminés de fibrinopeptides (Doolittle and Blombäck, 1964). L'avantage d'utiliser les séquences protéiques plutôt que des séquences nucléiques est que, du fait de la dégénérescence du code génétique et des substitutions synonymes des bases, les séquences protéiques sont plus conservées, ce qui les rend plus robustes aux phénomènes de convergence substitutionnelle, ou homoplasie. Elles permettent ainsi d'analyser des séquences ayant divergé il y a plus longtemps ou évoluant à un taux plus rapide. En outre, il semble que, d'après les observations faites jusqu'à maintenant, les séquences protéiques soient moins sensibles que les séquences nucléiques à des biais de composition, entrainant le regroupement de séquences partageant des compositions moléculaires similaires (voir sections 1.4.4 et 2.1). En revanche, les séquences protéiques deviennent non-informatives pour des degrés de divergence plus faible et ne permettent pas de résoudre des relations de parentés dont l'origine est récente.

Une distinction peut être faite entre les modèles de substitution mécanistes et empiriques. Les modèles mécanistes tentent de considérer explicitement les facteurs biologiques à l'oeuvre influençant le processus évolutif d'évolution des acides aminés. A l'aide de paramètres estimés à partir du jeu de données d'étude, ils ont, par exemple, pour but de quantifier les phénomènes de

biais mutationnels au niveau de l'ADN, de traduction des codons en acides aminés et du destin des mutations d'un point de vue de la sélection (Rodrigue et al., 2010). En revanche, les modèles empiriques décrivent les différents taux de substitutions entre acides aminés sans considérer explicitement les facteurs qui influencent le processus évolutif (Whelan and Goldman, 2001; Le and Gascuel, 2008). Les paramètres de ces modèles sont généralement estimés à partir d'un grand jeu de données issues de bases de données et fixés par la suite dans le but de les réutiliser sur d'autres jeux de données. Même s'ils sont moins élégants et qu'ils ne permettent pas d' appréhender les forces et les mécanismes évolutifs à l'oeuvre par rapport aux modèles mécanistes, les modèles empiriques semblent être plus efficaces lorsqu'il s'agit de reconstruire des arbres phylogénétiques (Yang, 2006). Il est à noter que les modèles mécanistiques ne sont pas l'apanage des séquences protéiques et qu'ils peuvent être construits afin de comprendre les mécanismes d'évolution moléculaire à l'oeuvre au niveau ADN ou codon (Yang and Nielsen, 2008).

La grande majorité des modèles de substitution protéiques utilisés en pratique sont des versions empiriques du modèle GTR, pour lesquels les échangeabilités et les fréquences d'équilibres ont été appris extérieurement au jeu de données analysé à partir d'une base de données d'alignements considérée comme représentative des séquences du vivant. La première matrice de substitution empirique a été construite par Dayhoff et al. (1978) à partir des familles de séquences protéiques disponibles à cette époque. À partir de ces familles et des arbres phylogénétiques correspondant, ils utilisèrent la reconstruction de séquences ancestrales par parcimonie (voir section 1.6.1) le long de ces arbres pour compiler les informations de substitutions observées entre paires d'acides aminés, afin de construire leur matrice de transitions. C'est à partir de cette matrice que toute une série d'autres matrices furent reconstruites selon la distance phylogénétique considérée. Dayhoff et al. (1978) ont approximé la matrice de probabilités de transitions pour une distance attendue de 0.01 substitution par site, produisant la matrice PAM1 (pour une "Point-Accepted Mutation") ou encore connue sous le nom de DAYHOFF. Ainsi, la matrice PAM250 est la matrice de transition acceptant 250 mutations pour 100 sites et est mieux adaptée pour retracer l'histoire évolutive de séquences éloignées. La matrice PAM fut par la suite améliorée à l'aide d'une approche différente d'apprentissage de ses paramètres et à partir d'une plus grande quantité de données pour produire les matrices BLOSUM62 (Henikoff and Henikoff, 1992) et JTT (Jones et al., 1992).

Les modèles proposés par la suite furent estimés en utilisant le Maximum de Vraisemblance (Adachi and Hasegawa, 1996) en considérant le modèle GTR de façon à estimer les échangeabilités et les fréquences d'équilibres du modèle (le Maximum de Vraisemblance est expliqué à la section 1.3). Cela implique l'estimation de 19x20/2 = 190 echangeabilités et 19 fréquences d'équilibres. Dans la pratique et lors du calcul de la vraisemblance sur un arbre, la matrice

$Q$ est normalisée afin d'obtenir des longueurs de branches interprétables. Ainsi, la contrainte $-sum_i \pi_i q_{ii} = 1$ est appliquée, de telle sorte que les distances évolutives soient mesurées en nombre moyen de changement par site. Ceci entraine l'élimination d'un paramètre à estimer, ce qui laisse 189 échangeabilités à estimer. Comme ces 208 (189 + 19) paramètres sont estimés au Maximum de Vraisemblance, un large jeu de données est nécessaire pour une optimisation correcte des valeurs d'échangeabilités et de fréquences d'équilibre. De nombreux modèles ont été constuits de cette façon. Nous citerons ici le modèle mtREV (ou MTMAM), spécifique des protéines mitochondriales de vertébrés ou mammifères (Adachi and Hasegawa, 1996), le modèle WAG (Whelan and Goldman, 2001), appris à partir de données nucléaires et enfin le modèle LG (Le and Gascuel, 2008), modèle le plus performant à l'heure actuelle. LG est une amélioration du modèle WAG en prenant en compte plus de données et en prenant en compte la variation de vitesse d'évolution entre sites (voir section 1.4.1). La Figure 1.5 montre les matrices symétriques d'échangeabilités des modèles WAG et LG.

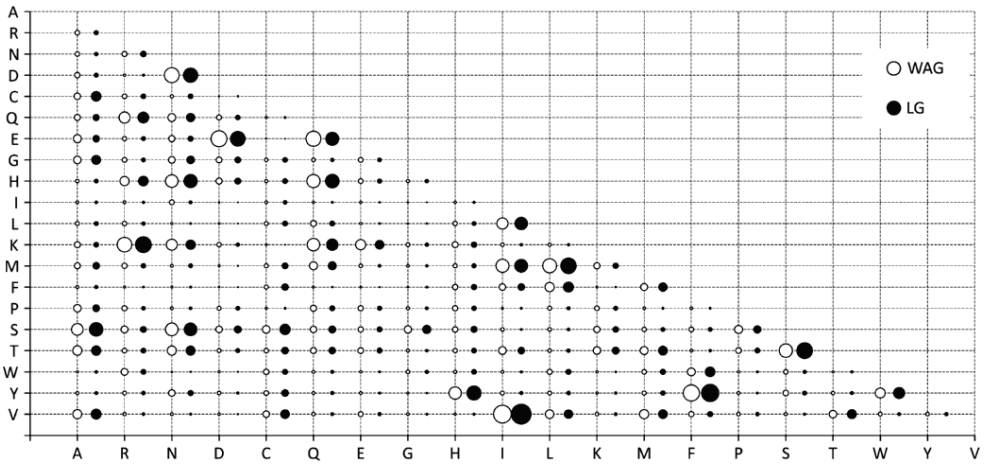



FIGURE 1.5 – Comparaison des échangeabilités des modèles WAG et LG. Cette figure a été extraite de la figure 2 de Le and Gascuel (2008)

Il est possible de voir dans la figure 1.5 que les acides aminés ayant des propriétés biochimiques similaires ont de plus grandes échangeabilités et vont avoir tendance à s'échanger

plus fréquemment au cours de l'évolution. Par exemple, les acides aminés isoleucine (I) et valine (V), ainsi que les acides aminés arginine (R) et lysine (K) ont de fortes échangeabilités. En revanche, la cystéine ou la proline, qui sont deux acides aminés très particuliers (du fait de la présence d'un groupement thiol (-SH) pour la cystéine, permettant notamment de faire des ponts disulfures et du fait de la présence d'un cycle pour la proline, impactant fortement son encombrement stérique), ont des échangeabilités très faibles avec l'ensemble des autres acides aminés. L'autre caractéristique importante que ces matrices révèlent est la plus grande échangeabilité entre acides aminés codés par des codons séparés par une seule mutation, et inversement dans le cas d'acides aminés séparés par deux à trois mutations. Par exemple, l'acide aspartique (D) et l'acide glutamique (E) peuvent se substituer par une unique mutation de base au niveau de leur codon. À l'opposé, les codons codant pour la glycine (G) et la leucine (L) sont séparés par 2 mutations et ont des échangeabilités quasi nulles.

Lorsque de tels modèles empiriques sont utilisés dans la littérature, le profil du modèle est souvent modifié pour être ajusté aux fréquences observées de l'alignement (Cao et al., 1994). Ces modèles ont alors un suffixe '+F' ou '-F'. Comme la plupart du temps le modèle de substitution utilisé est stationnaire et que l'on suppose que les séquences sont à l'équilibre compositionnel, il devient justifier d'utiliser ces fréquences observées comme fréquences d'équilibres du modèle, tout en gardant les échangeabilités empiriques propres au modèle. Généralement, fixer le profil aux fréquences observées améliore l'ajustement du modèle aux données, mais pas nécessairement (Groussin et al., 2013a).

## 1.3 Le maximum de vraisemblance en phylogénie

Le maximum de vraisemblance ou Maximum Likelihood (ML) est une méthode statistique largement utilisée pour estimer les paramètres d'une distribution de probabilité d'un échantillon donné. Son application en phylogénie s'explique par la volonté de prédire quel les paramètres du scénario évolutif ayant généré les données observées actuellement. Felsenstein (1981) fut le premier à développer un algorithme efficace du calcul de la vraisemblance en phylogénie, permettant au maximum de vraisemblance d'être progressivement considéré lors d'analyses phylogénétiques jusqu'à devenir au cours des années 1990 la méthode de référence pour les phylogénéticiens.

### 1.3.1 Principe général de la vraisemblance

- Soit $\mathbf{x} = (x_1, x_2, x_3, \ldots, x_\ell)$ un échantillon de valeurs d'une variable aléatoire $X$ provenant d'une distribution de paramétre(s) $\theta$ inconnu(s) :

– Si l'échantillon est aléatoire, alors les $x_i$ ($1 \leq i \leq \ell$) sont indépendantes et indentiquement distribuées et chacune a une probabilité $\mathbb{P}(x_i|\theta)$.

– Si l'ensemble de l'échantillon est considéré, alors la *fonction de vraisemblance* $L(\theta)$ associée est égale à la probabilité simultanée :

$$\begin{aligned} L(\theta) = \mathbb{P}(\mathbf{x}|\theta) &= \mathbb{P}(x_1|\theta) \times \mathbb{P}(x_2|\theta) \times \cdots \times \mathbb{P}(x_\ell|\theta) \\ &= \prod_{i=1}^{\ell} \mathbb{P}(x_i|\theta) \end{aligned}$$

Ainsi, la vraisemblance est égale à la *probabilité des données ($x$) sachant les paramètres du modèle ($\theta$)*.

– La plupart des méthodes n'utilisent pas directement la fonction de vraisemblance, mais plutôt son logarithme, afin d'éviter des problèmes numériques liés à la manipulation de valeurs potentiellement très petites :

$$\ln L(\theta) = \sum_{i=1}^{\ell} \ln \mathbb{P}(x_i|\theta)$$

### 1.3.2 Application à la phylogénie

Transposons à présent cette notion de vraisemblance à la problématique phylogénétique. Il est à noter que l'ensemble des principes développés ici est transposable à n'importe quel alphabet moléculaire.

- Soit un alignement $D$ comprenant un nombre $d$ de sites :

  – Chaque site dans l'alignement est désigné par le terme $C^{(i)}$ ($1 \leq i \leq d$).

  – Dans ce cas, l'expression de la vraisemblance associée au site $i$ est donnée par :

  $$L^{(i)} = \mathbb{P}(C^{(i)}|\tau, \mathbf{b}, \vartheta)$$

  avec $\tau$, la topologie considérée, $\mathbf{b}$ le vecteur des longueurs de branches, et $\vartheta$ les paramètres du modèle d'évolution utilisé. Par la suite et par souci de simplification des notations, nous regrouperons l'ensemble des paramètres phylogénétiques $\tau, \mathbf{b}$ et $\vartheta$ sous la variable $\theta$.

  – On en déduit l'expression de la vraisemblance pour l'ensemble des sites indépen-

dants :

$$L = \mathbb{P}(D|\theta) = \prod_{i=1}^{d} L^{(i)} = \prod_{i=1}^{d} \mathbb{P}(C^{(i)}|\theta)$$

- La question est maintenant de savoir comment calculer la vraisemblance pour un site (ou colonne) donné $C^{(i)}$ de l'alignement. Considérons l'arbre phylogénétique de la figure 1.6. Cet arbre est un arbre à quatre feuilles de topologie $\tau$ et dont les longueurs de branches $\mathbf{b}$ sont fixées.

  - $S_1$, $S_2$, $S_3$ et $S_4$ représentent les feuilles de l'arbre.
  - $N_1$, $N_2$ et $N_3$ représentent les nœuds internes.
  - Les états de caractères correspondants sont dénotés par $s_1$, $s_2$, $s_3$, $s_4$, $n_1$, $n_2$, $n_3 \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ pour le cas nucléique.

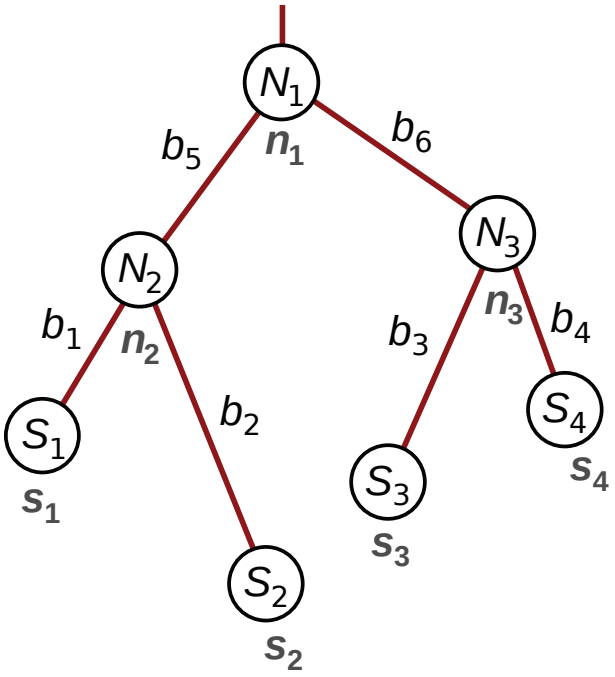



FIGURE 1.6 – Exemple d'arbre phylogénétique à quatre feuilles.

- Pour calculer la vraisemblance à un site $i$ de l'alignement on prend en compte tous les

scénarios évolutifs possibles à chaque nœud interne de l'arbre. Ainsi :

$$L^{(i)} = \sum_{n_1}\sum_{n_2}\sum_{n_3}\mathbb{P}(n_1)\mathbb{P}(n_2|n_1,b_5)\mathbb{P}(n_3|n_1,b_6)\mathbb{P}(s_1|n_2,b_1) \times \mathbb{P}(s_2|n_2,b_2)\mathbb{P}(s_3|n_3,b_3)\mathbb{P}(s_4|n_3,b_4)$$

Soit :

$$L^{(i)} = \sum_{n_1}\mathbb{P}(n_1)\{\sum_{n_2}(\mathbb{P}(n_2|n_1,b_5)\mathbb{P}(s_2|n_2,b_2)\mathbb{P}(s_1|n_2,b_1)) \times \sum_{n_3}(\mathbb{P}(n_3|n_1,b_6)\mathbb{P}(s_3|n_3,b_3)\mathbb{P}(s_4|n_3,b_4))\}$$

- Sous l'hypothèse que le processus markovien modélisant l'évolution des séquences a atteint *l'état stationnaire*, on a :

$$\mathbb{P}(n_1) = \pi_{n_1}$$

avec $\pi_{n_1}$, la fréquence de l'état de caractère $n_1$.

La vraisemblance d'une colonne $C^{(i)}$, sachant les paramètres $\theta$ du modèle, est alors vue comme l'intégration probabiliste de tous les scénarios substitutionnels possibles le long de l'arbre phylogénétique ayant abouti aux données observées aux feuilles.

### 1.3.3 Maximisation de la vraisemblance

L'approche du Maximum de Vraisemblance, par la suite nommée ML pour Maximum Likelihood, consiste à maximiser la fonction de vraisemblance. Par cette maximisation, il s'agit de déterminer l'ensemble des paramètres $\tau, \mathbf{b}$ et $\vartheta$ pour lesquels les données sont les plus probables. Cette fonction est une fonction de plusieurs variables qu'il faut tenter de maximiser. Par exemple, si l'on se place dans le cas homogène et que l'on considère que toutes les branches sont caractérisées par le même modèle de substitution, lorsque de nouveaux paramètres de ce modèle sont proposés, les probabilités de transition exprimées dans la formulation de $L^{(i)}$ sont modifiées, de telle sorte que $L^{(i)}$ et donc $L$ sont également modifiées, permettant d'estimer quels sont les paramètres du modèle qui globalement à l'échelle de l'alignement maximisent la vraisemblance.

Cette fonction de vraisemblance est une fonction à plusieurs variables dont la recherche du maximum global est complexe. Premièrement, cette fonction est dépendante de paramètres discrets (la topologie) et continus (longueurs de branches, paramètres du modèle de substitution), ce qui entraine que l'estimation du maximum ne peut se faire de manière jointe entre tous les paramètres. Deuxièmement, la combinatoire des valeurs de paramètres possibles est gigantesque, entrainant la présence de maxima locaux dans l'espace des paramètres rendant difficile la détermination du maximum global. Enfin, il n'est pas toujours possible d'avoir accès aux dérivées

analytiques premières et secondes de la fonction de vraisemblance pour tous les paramètres. Toutes ces raisons expliquent pourquoi cette recherche du maximum global de la vraisemblance s'effectue en pratique par des approches *heuristiques* à travers l'optimisation numérique des paramètres. La description de toutes ces approches mathématiques dépasse le cadre de cette thèse. Il est toutefois possible de mentionner les approches de Brent (Brent et al., 1973) (méthode très utilisée pour les paramètres pour lesquels les dérivées de la fonction de vraisemblance ne sont pas connues), les méthodes de Newton ou Newton-Raphson (Felsenstein et al., 1996; Yang, 2000) utilisant les dérivées ou encore les méthodes numérique dites de "hill-climbing" comme le gradient conjugué ou la méthode Broyden-Fletcher-Goldfarb-Shanno (BFGS).

### 1.3.4 Algorithme d'élagage (Pruning algorithm)

Felsenstein (1981) a décrit un algorithme de programmation dynamique permettant de calculer efficacement la fonction de vraisemblance présentée plus haut. Cet algorithme, dit du "pruning" ou d'élagage utilise la structure de l'arbre et la possibilité de reformuler la vraisemblance sous forme de produits de vraisemblances conditionnelles des états aux noeuds. Considérons la figure 1.7, qui représente le cas général d'un noeud de l'arbre, en reprenant les notations de la figure 1.6 :

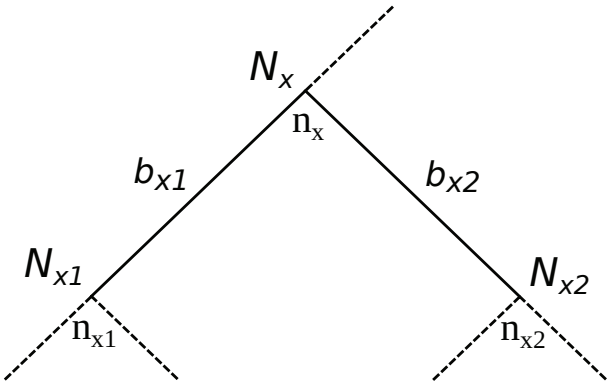



FIGURE 1.7 – Cas général d'un noeud interne.

La vraisemblance conditionnelle du noeud $N_x$ pour l'état $n_x$ peut alors s'exprimer de la façon suivante :

$$L^{(i)}_{N_x}(n_x) = \begin{cases} 1 \text{ si } N_x \text{ est une feuille avec l'état} n_x \\ 0 \text{ si } N_x \text{ est une feuille avec un état différent de } n_x \\ \sum_{n_{x1}} \mathbb{P}(n_x|n_{x1}, b_{x1}) L^{(i)}_{N_{x1}}(n_{x1}) \times \sum_{n_{x2}} \mathbb{P}(n_x|n_{x2}, b_{x2}) L^{(i)}_{N_{x2}}(n_{x2}) \text{ sinon.} \end{cases}$$

En partant des feuilles et en remontant dans l'arbre, il est possible de calculer les vraisemblances contionnelles pour chaque état, qui représentent, pour un noeud donné, la probabilité d'observer les données aux feuilles sous ce noeud sachant l'état au noeud. Le calcul des vraisemblances conditionnelles aux noeuds s'effectue donc *récursivement* en remontant dans l'arbre : pour un noeud donné, les vraisemblances conditionnelles ne sont calculées que lorsque celles des noeuds descendants l'ont été. L'algorithme de pruning permet alors de calculer successivement les probabilités de beaucoup de sous-arbres, ce qui est très efficace notamment lors de l'exploration des topologies, évitant ainsi d'avoir à recalculer des vraisemblances pour les sous-arbres n'ayant pas subi de changements topologiques.

### 1.3.5 Brève introduction au bayesien

L'ensemble des modèles développés dans cette thèse, ainsi que la grande majorité des calculs phylogénétiques ont été effectué dans le cadre du ML. C'est pourquoi l'approche bayesienne n'est ici que partiellement introduite. Pour une explication plus précise de l'utilisation du bayesien en phylogénie, se référer aux livres de Felsenstein (2004) et Yang (2006).

Tout d'abord, il faut se rendre compte que le ML produit un résultat qui est une unique réalisation du modèle phylogénétique. C'est la réalisation qui maximise la vraisemblance d'observer les données sachant ce modèle. Dans le cadre bayesien, le résultat est une *distribution* de réalisations. Les méthodes bayesiennes utilisent la notion de vraisemblance des données mais également la notion de probabilité *a priori* des réalisations ou des paramètres du modèle $\theta$. Ainsi, d'après le théorème de Bayes,

$$\mathbb{P}(D|\theta) = \frac{\mathbb{P}(D, \theta)}{\mathbb{P}(\theta)}$$

soit :

$$\mathbb{P}(D|\theta) = \frac{\mathbb{P}(D)\mathbb{P}(\theta|D)}{\mathbb{P}(\theta)}$$

d'où :

$$\mathbb{P}(\theta|D) = \frac{\mathbb{P}(\theta)\mathbb{P}(D|\theta)}{\mathbb{P}(D)}$$

La probabilité $\mathbb{P}(\theta|D)$ est la probabilité *a posteriori* des paramètres du modèle sachant les données. $\mathbb{P}(D)$ est la vraisemblance marginale des données et est appelé le facteur de Bayes, tandis que $\mathbb{P}(\theta)$ représente la distribution *a priori* des paramètres et $\mathbb{P}(D|\theta)$ la vraisemblance des données.

En pratique, le calcul des probabilités postérieures des arbres phylogénétiques ou des paramètres du modèle de substitution est analytiquement impossible. C'est pourquoi des méthodes numériques de type chaines de Markov avec technique de Monte Carlo (MCMC pour "Markov Chain Monte Carlo") ont été développées et utilisées (Yang and Rannala, 1997) pour échantillonner efficacement dans l'espace des valeurs de paramètres et topologies postérieures. Le principe des MCMC repose sur l'idée qu'une chaine de Markov prenant la forme d'une marche *guidée* à travers l'espace multidimensionnel des paramètres peut être utilisée pour approcher la distribution de probabilité de ces paramètres en échantillonnant les valeurs de manière périodique. En phylogénie, chaque pas de la chaine MCMC va modifier aléatoirement les paramètres du modèle et donc la vraisemblance et permettre d'explorer les combinaisons de paramètres autour de la vraisemblance maximale. L'algorithme le plus fréquemment utilisé permettant de *guider* le MCMC est l'algorithme de Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970).

## 1.4 Vers des modèles plus réalistes d'un point de vue biologique

Les modèles de substitutions présentés dans les sections précédentes sont très simplifiés par rapport à la complexité des mécanismes évolutifs agissant au niveau des séquences biologiques. Malgré la démonstration de l'intérêt de leur utilisation pour comprendre certains aspects de l'histoire évolutive des gènes, ils peuvent devenir limités voire systématiquement biaisés du fait d'hypothèses trop simplificatrices quant aux processus évolutifs à l'oeuvre (Philippe and Roure, 2011; Anisimova et al., 2013). Cela suggère le développement de modèles plus réalistes d'un point de vue biologiques et qui sont inspirés de connaissances *a priori* sur les patrons moléculaires. Cette partie présente succintement quelques détails concernant ces modèles plus complexes. De très nombreux modèles complexifiant le modèle GTR simple ont été présentés par le passé et une revue complète de la littérature concernant ces modèles dépasse le cadre de cette thèse. En revanche, les modèles hétérogènes en sites (sections 1.4.1 et 1.4.3) et en temps (sections 1.4.2 et 1.4.4) sont abordés, quoique de manière introductive seulement. Ils sont plus largement détaillés dans les introductions des deux premiers manuscrits (sections 2.1 et 2.2), présentant le développement de nouveaux modèles hétérogènes en temps et en sites.

### 1.4.1 Variations des taux d'évolution entre sites

Les bases ou acides aminés n'évoluent pas tous à la même vitesse le long des séquences d'ADN ou de protéines. De multiples raisons peuvent expliquer cela, allant de mécanismes neutres aux mécanismes sélectifs. Il a été plusieurs fois démontré qu'ignorer ces variations de taux d'évolution peut avoir de lourdes conséquences sur les analyses phylogénétiques (Tateno et al., 1994; Yang, 1996a; Sullivan et al., 2001). Pour modéliser ces variations de taux entre sites dans le cadre d'un modèle Markovien de substitution en ML, plusieurs approches ont été proposées (Yang, 2006). La plus utilisée considère qu'il n'y a pas de connaissance *a priori* des taux spécifiques des sites et emploie une distribution $\Gamma$ discrète pour fournir des taux possibles aux sites. Ce modèle, proposé par Yang (1994), permet à un site d'être modélisé par un mélange de matrices qui ne diffèrent que par leur taux. La matrice de substitution reste identique parmi les composantes du modèle de mélange, à un facteur multiplicatif prêt. Pour un site donné, $K$ vraisemblances vont être calculées avec ces $K$ modèles dans le cas d'une distribution discrétisée en $K$ catégories. Ainsi, la vraisemblance d'un site est la vraisemblance pondérée des vraisemblances calculées avec chaque taux, en considérant que les composantes de la distribution discrète sont équiprobables. La vraisemblance totale devient :

$$L = \mathbb{P}(D|\theta) = \prod_{i=1}^{d} L^{(i)} = \prod_{i=1}^{d} \left\{ \sum_{k=1}^{K} \frac{1}{K} \mathbb{P}(C^{(i)}|\theta_k) \right\}$$

Le paramètre d'échelle $\beta$ de la distribution $\Gamma$ est fixé arbitrairement à la valeur du paramètre de forme $\alpha$, de telle sorte que la moyenne de la distribution soit égale à 1 et que le taux global de la matrice de substitution soit aussi de 1 pour exprimer les longueurs de branches en nombre de substitutions moyen par site (voir les explications dans les livres de Felsenstein (2004) et Yang (2006)). Ce paramètre $\alpha$ est estimé au ML et permet de définir des taux différents pour les différentes catégories (voir figure 1.8).

### 1.4.2 Variations des taux d'évolution dans le temps

L'utilisation d'une distribution $\Gamma$ pour modéliser la variation de taux entre sites ne permet pas de prendre en compte les variations potentielles des taux dans le temps. En effet, pour une catégorie donnée de la distribution $\Gamma$ discrète, la vraisemblance est calculée le long de l'arbre en considérant que le taux spécifique de cette catégorie est le même entre lignées. Encore une fois, de multiples raisons peuvent expliquer pourquoi certaines lignées se mettent à évoluer plus lentement ou plus rapidement dans le temps, comme lors d'adaptations à des environnements changeants. Les modèles permettant de relâcher l'hypothèse de constance des taux dans le temps sont appelés modèle de *covarion* (Tuffley and Steel, 1998; Huelsenbeck et al., 2002). Un exemple d'un

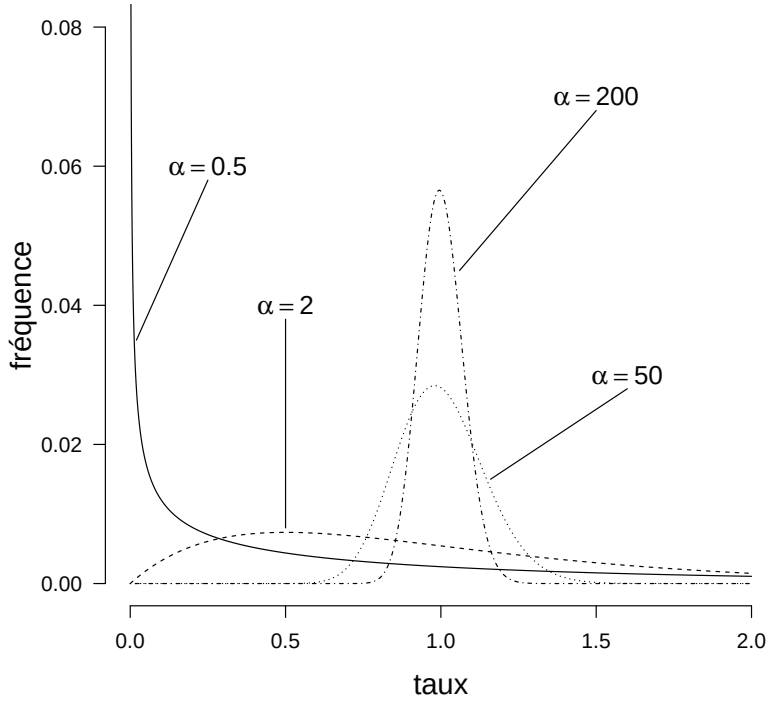



FIGURE 1.8 – Distributions Γ, selon la valeur du paramètre de forme $\alpha$.

tel modèle est le modèle de Galtier (2001), qui est en fait une extension du modèle Γ pour la variation des taux entre sites. Ainsi, un site peut dans le temps sauter d'une classe de vitesse de la distribution Γ à une autre, pour être caractérisé par une vitesse rapide dans certaines lignées ou par une vitesse lente dans d'autres. Galtier (2001) a montré que ne pas prendre en compte ces phénomènes aux niveaux des molécules d'ARN ribosomiques pouvait entrainer une mauvaise estimation de la variance de la distribution des taux entre sites, ainsi qu'une moindre capacité à détecter les substitutions multiples.

### 1.4.3 Variations des processus évolutifs entre sites

La plupart des modèles de substitutions décrits dans la littérature et considérés en pratique ne contiennent qu'une seule matrice de taux de substitutions, utilisée pour calculer la vraisemblance de tous les sites. Cependant, l'observation des profils observés de sites indiquent que le processus évolutif est hétérogène entre sites. Ceci reflète les différentes pressions de sélection s'exerçant le long d'une protéine, selon l'implication des sites dans des structures secondaires ou fonctions particulières. Des modèles de substitutions prenant en compte ce phénomène ont été développés et ont montré leur efficacité vis à vis de l'ajustement aux données ou de leur plus grande résistance au biais d'attraction des longues branches (voir Introduction du manuscrit présenté en

section 2.2). Par exemple, le modèle CAT (Lartillot and Philippe, 2004) permet de proposer un grand nombre de modèles de substitution et d'allouer ces modèles aux sites, permettant à un site d'être modélisé par un modèle qui lui est très spécifique. Ces modèles de substitution varient par leur profil, qui peuvent être spécifiques de 2 à 4 acides aminés seulement afin de refléter les profils de site observés dans les séquences réelles. Une autre possibilité est d'utiliser un modèle de mélange dans le même esprit que lorsque la variation des taux entre sites est prise en compte avec la distribution $\Gamma$. Un site est alors modélisé par un mélange de matrices de substitution qui ont des échangeabilités et/ou des profils différents, et la vraisemblance totale du site est la moyenne pondérée des vraisemblances calculées avec chacune des matrices. La vraisemblance totale devient alors :

$$L = \mathbb{P}(D|\theta) = \prod_{i=1}^{d} L^{(i)} = \prod_{i=1}^{d} \left\{ \sum_{z=1}^{Z} \sigma_z \sum_{k=1}^{K} \frac{1}{K} \mathbb{P}(C^{(i)}|\theta_{k,z}) \right\}$$

- avec $Z$ le nombre de composantes (matrices de substitution) du mélange
- et $\sigma_z$ le poids spécifique de la composante $z$, sachant que :

$$\sum_{z=1}^{Z} \sigma_z = 1.$$

Lorsqu'un site est à la fois modélisé par un mélange de taux à $K$ composantes et un mélange de modèles à $Z$ composantes, $K \times Z$ vraisemblances sont calculées par site. Cela entraine un coût certain, à la fois en temps de calcul et en mémoire. Malgré cela, la modélisation des deux phénomènes hétérogènes (taux et processus) est nécessaire pour que le modèle s'ajuste correctement aux données. Le et al. (2012) ont alors proposé deux modèles, LG4M et LG4X, qui présentent l'avantage de prendre en compte ces hétérogénéités tout en réduisant la quantité de calcul. Ces modèles supposent que le processus varie selon le taux d'évolution. En d'autres termes, chaque catégorie de la distribution $\Gamma$ possède maintenant son propre processus, réduisant le nombre de vraisemblances à calculer par sites. Ils ont montré que, comparativement aux modèles site-homogènes, de grands gains de vraisemblance étaient observés, montrant l'intérêt de les utiliser lors d'analyses phylogénétiques.

Enfin, si des connaissances *a priori* sont disponibles sur une variation des taux ou des processus entre sites ou gènes, il est possible d'utiliser un modèle qui combine les différentes partitions. Ces partitions ont leur propre taux ou processus. Par exemple, il est connu que les trois positions des codons n'évoluent pas aux mêmes taux car selon l'endroit où une mutation apparait dans le codon, l'impact sur l'acide aminé peut être très différent. Il peut alors devenir intéressant de

créer une partition par positions de codons et d'utiliser un modèle qui combine les partitions. De tels modèles ont été proposés par Yang (1995, 1996b) et sont efficaces en pratique (Shapiro et al., 2006). Récemment, Zoller and Schneider (2012) ont proposé un modèle améliorant le modèle simple ne contenant qu'une seule matrice empirique. Ce modèle, appelé PCMA, est un modèle semi-empirique. Zoller et Schneider ont estimé environ 3,000 matrices d'échangeabilités à partir d'autant d'alignements extraits d'une large base de données. En utilisant ensuite une Analyse en Composantes Principales (ACP ou PCA en anglais), ils ont pu extraire les 190 axes majeurs de variabilité des échangeabilités entre gènes (190 car il y a 190 échangeabilités dans la matrice $S$). Le modèle utilisé par la suite pour calculer la vraisemblance de l'alignement ne contient qu'une seule matrice, dont la matrice d'échangeabilités est estimée à partir des données en utilisant les axes de variabilité de l'ACP. Si les $q$ premiers axes de l'ACP représentant la plus grande variance sont considérés, le modèle estime $q$ positions le long de ces $q$ axes et calcule la matrice d'échangeabilité correspondant à ces positions en renversant l'ACP. Cette manipulation permet d'explorer l'espace des échangeabilités possibles dans un espace de faible dimension donné par l'ACP. Cette technique se rapproche fortement de celle que nous avons mise en oeuvre dans le modèle COaLA (Groussin et al., 2013a) présenté dans la section 2.1, qui permet d'estimer efficacement les fréquences d' équilibres d'une branche donnée avec un modèle hétérogène entre lignées. À noter que le modèle PCMA ne rend pas compte de l'hétérogénéité du processus entre sites, mais de l'hétérogénéité du processus entre gènes. Ils ont montré que leur modèle était très efficace et capable d'atteindre des performances égales à celles des meilleurs modèles de mélanges (UL3 (Le et al., 2008b), hétérogène en sites).

### 1.4.4 Variations des processus évolutifs dans le temps

Les compositions moléculaires en bases des génomes ou en acides aminés des protéomes peuvent varier de manière significative d'une espèce à l'autre. Par exemple, dès 1962, Sueoka a mis en évidence que les taux de G+C génomiques étaient très hétérogènes entre espèces bactériennes, pouvant aller de 25% à 75% (Sueoka, 1962; Bentley and Parkhill, 2004). Plusieurs théories ont été développées afin de tenter d'expliquer ces observations. Il a ainsi été proposé que la sélection pouvait agir au niveau du génome afin de sélectionner les compositions qui apporteraient un avantage sélectif à l'organisme d'un point de vue écologique ou physiologique. Par exemple, il existe un lien entre mode de vie aérobie ou anaérobie et composition en GC, de telle sorte que la vie aérobie serait responsable de l'augmentation du taux de GC des espèces aérobiques (Naya et al., 2002). Bien que cette corrélation soit effective, aucun mécanisme clair expliquant le lien vie aérobie et augmentation du G+C n'a été proposé. Des théories neutralistes ont également été mises en avant, suggérant que ces différences de GC génomiques étaient essentiellement

dûes à des biais mutationnels (Gautier, 2000; Rocha et al., 2006). Cependant, deux études récentes publiées conjointement ont montré que le bias mutationnel vers AT était universel chez les procaryotes, même chez les espèces ayant un fort taux de GC génomique (Hershberg and Petrov, 2010; Hildebrand et al., 2010). Cela suggère qu'une force contrecarre l'effet du biais mutationnel. La question est maintenant de savoir si ce sont bien des processus sélectifs qui sont responsables de cet effet, ou si d'autres processus neutres, tels que le biais de conversion génique biaisé vers GC sont à l'oeuvre (Galtier and Duret, 2007; Hershberg and Petrov, 2010; Hildebrand et al., 2010).

Les variations de taux de GC génomique entre espèces procaryotes se répercutent en grande partie sur la composition en acides aminés des protéines. L'hétérogénéité de composition en GC génomique est même le facteur majeur expliquant la variance observée des compositions protéomiques (Boussau et al., 2008). Ceci s'explique par le fait que certains acides aminés possèdent des codons plus riches en GC que d'autres, entrainant l'enrichissement de ces acides aminés chez les espèces riches en GC. En outre, de nombreux facteurs environnementaux et adaptatifs peuvent en partie expliquer les variations de compositions en acides aminés. Ainsi, il a été montré que la température (Singer and Hickey, 2003; Lobry and Necsulea, 2006; Tekaia and Yeramian, 2006; Zeldovich et al., 2007; Boussau et al., 2008) ou la salinité (Madern et al., 2000; Kiraga et al., 2007; Paul et al., 2008) de vie de l'organisme influençait grandement les compositions des protéines. Il est souvent considéré que ces variations compositionnelles sont en lien avec une conservation ou une augmentation de la stabilité des protéines codées par ces organismes résidant dans des environnements extrêmes.

Malgré ces nombreuses observations d'hétérogénéités de compositions, la plupart des modèles de substitutions utilisés dans la littérature font l'hypothèse que les compositions ne changent pas au cours du temps et que les espèces évoluent à l'équilibre compositionnel. Il a plusieurs fois été montré que l'utilisation de ces modèles sur des données présentant des variations de composition pouvait entraîner de mauvaises inférences phylogénétiques (Foster and Hickey, 1999; Foster, 2004; Jermiin et al., 2004), notamment en regroupant les espèces partageant les mêmes compositions. Récemment, un modèle sophistiqué permettant de modéliser l'évolution de cette hétérogénéité de composition dans le temps a été proposé dans le cadre bayesien (Blanquart and Lartillot, 2006, 2008). Une plus grande description de tous ces modèles est effectuée dans l'introduction du manuscrit présenté dans la section 2.1.

## 1.5 Nombre de paramètres d'un modèle et conséquences

En modélisation, une question importante est de savoir à quel point un modèle donné est capable de s'ajuster aux données étudiées. Plus précisemment, cet ajustement mesure l'écart entre les

données observées et les données attendues selon le modèle.

### 1.5.1 Notion de biais et de variance d'un modèle

Comme vu précédemment, un modèle contient des paramètres dont les valeurs peuvent être estimées à partir des données. Un modèle peut alors être plus ou moins complexe selon le nombre de paramètres considérés, qui visent à expliquer les phénomènes biologiques ayant abouti à la distribution des données observées. À première vue, plus le modèle est complexe, plus il a des chances d'expliquer précisemment les données. Seulement, le modèle est aussi susceptible de générer des erreurs, dont l'expression est Erreur = Biais + Variance. Prenons l'exemple d'une variable X dont la distribution des valeurs (points rouges) est représentée dans la figure 1.9. Si l'on essaye d'expliquer la distribution de ces valeurs observées selon un modèle polynômial, ce modèle peut être plus ou moins complexe selon le degré du polynome, les coefficients du polynôme étant les paramètres du modèle.

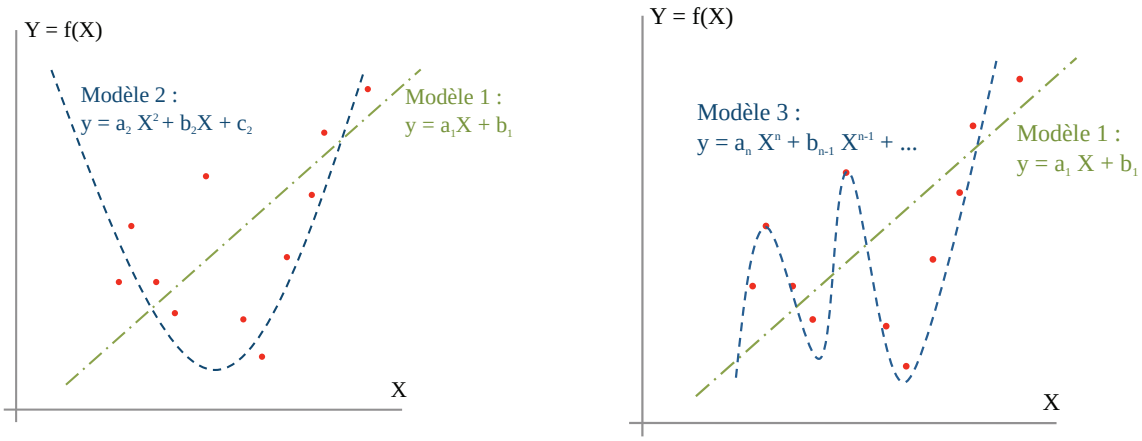



FIGURE 1.9 – Exemple d'un modèle polynomial et ajustement aux données

Ce que l'on voit, c'est que le modèle 1 est biaisé et imprécis. Sa faible complexité ne lui permet pas de s'ajuster correctement aux données. Un modèle est à la fois un mécanisme descriptif permettant d'expliquer les caractéristiques des données observées et la fois un mécanisme qui doit être capable de les reproduire. Ainsi, malgré le fait que le modèle 3 "colle" aux données, il perd sa capacité de prédiction (on dit qu'il a une plus grande variance). En effet, si on rajoute des données observées dans le graphe de droite, il y a de fortes chances pour que ces données soient éloignées de la fonction polynomiale du modèle 3. À ce stade, il s'agit donc de trouver un compromis entre biais et variance (Figure 1.10) de telle sorte que le modèle ait un minimum de biais et de variance. Dans cet exemple, le modèle 2 représente ce compromis.

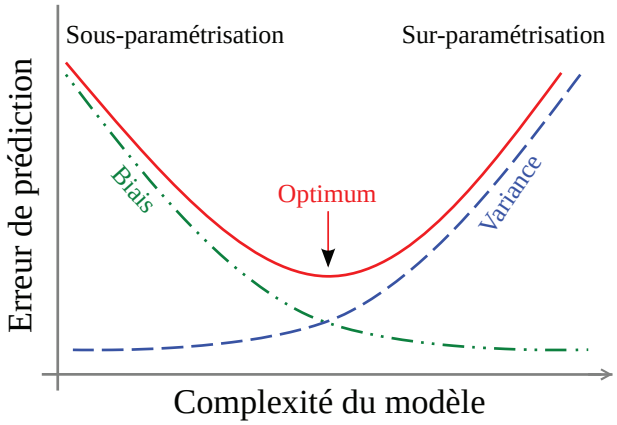



FIGURE 1.10 – Ajustement aux données et compromis entre biais et variance d'un modèle

De manière générale, il s'agit d'identifier le meilleur modèle en terme de complexité optimale, c'est à dire en terme de nombre de paramètres. En phylogénie, les modèles les plus complexes vont dans la grande majorité des cas aboutir à une meilleure vraisemblance. Cependant, ce gain de vraisemblance peut être dû à un trop fort ajustement aux données et donc augmenter la variance. Comparer deux ou plusieurs modèles passe par la comparaison des vraisemblances maximales à travers des critères statistiques abordés ci-dessous.

### 1.5.2 Comparaison de modèles emboités

Les modèles emboités sont des modèles qui sont des cas particuliers d'autres modèles. La grande majorité des modèles nucléiques sont emboités et représentent des cas particuliers du modèle le plus général, le modèle GTR. La figure 1.11 montre cet emboitement en précisant quelles sont les hypothèses faites sur les paramètres pour qu'un modèle donné soit un cas particulier d'un modèle plus général.

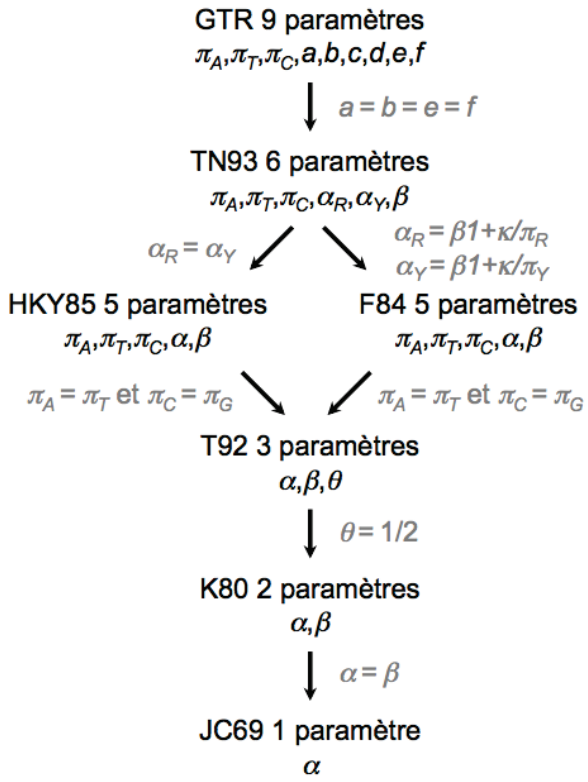



FIGURE 1.11 – Emboitement des modèles nucléiques. Cette figure a été extraite du livre de Perrière and Brochier-Armanet (2010) (figure 2.6)

Un des tests les plus utilisés pour la sélection de modèles est le test LRT, pour Likelihood Ratio Test, qui permet de comparer des modèles emboités.

- Soit $\mathcal{M}_0$ ($n_0$ paramètres) et $\mathcal{M}_1$ ($n_1 > n_0$ paramètres) deux modèles avec $\mathcal{M}_0$ emboité dans $\mathcal{M}_1$.
- Après calcul des deux vraisemblances $\mathcal{L}_0$ et $\mathcal{L}_1$ associées aux modèles $\mathcal{M}_0$ et $\mathcal{M}_1$, la statistique LRT peut-être calculée comme suit :

$$LRT = 2 \times ln(\frac{\mathcal{L}_1}{\mathcal{L}_0})$$

- Il peut-être montré que le LRT suit une loi de $\chi^2$ avec $n_1$ - $n_0$ degrés de liberté.
- Le LRT quantifie l'augmentation attendue de log-vraisemblance en passant du modèle $\mathcal{M}_0$ au modèle $\mathcal{M}_1$. Un jeu de données montrant un excès d'augmentation entraîne le rejet du modèle $\mathcal{M}_0$.

### 1.5.3 Comparaison de modèles non-emboités

Dans de nombreux cas, comme avec les modèles protéiques, les modèles comparés ne sont pas emboités. Bien que l'utilisation du LRT ait été étudiée pour la comparaison de modèles non-emboités, cette approche ne s'est pas répandue dans la littérature (Yang, 2006; Lewis et al., 2011), la raison principale étant la difficulté à déterminer explicitement la formule de la distribution de la statistique LRT sous l'hypothèse nulle qu'un des deux modèles est vrai. Le critère AIC, pour Akaike Information Criterion permet lui de comparer des modèles non-emboités en calculant pour chaque modèle le score suivant :

$$AIC = -2 \times \ln L + 2 \times K$$

avec $\ln L$ la vraisemblance maximale du modèle et $K$ le nombre de paramètres du modèle. Ainsi, le modèle ayant le plus faible score d'AIC est préféré. Une variante de ce score AIC a été proposé, qui représente en fait une correction du score AIC pour les jeux de données de petite taille. Ce score, nommé AICc, se calcule comme suit :

$$AICc = AIC + \frac{2 \times K \times (K+1)}{d - K - 1}$$

avec $d$ la taille de l'alignement, c'est à dire le nombre de sites.

Un autre critère très souvent utilisé est le BIC, pour Bayesian Information Criterion. Il est égal à :

$$BIC = -2 \times \ln L + K \times \ln d$$

On peut remarquer tout d'abord que dans la condition où $d >> K$, le critère AICc converge vers le critère AIC. Ensuite, ces critères ne s'appliquent pas de la même façon selon les jeux de données. En effet, le critère BIC pénalise beaucoup plus les modèles riches en paramètres que ne le fait le critère AIC, de telle sorte qu'il n'est pas vraiment recommandé de l'utiliser sur de petits jeux de données. En revanche il est plus adapté aux cas où le jeu de données est grand (comme dans le cas de concaténats en phylogénie) afin de pénaliser plus fortement les modèles riches en paramètres. Finalement, tous ces critères (ainsi que le LRT) permettent d'autoriser l'ajout de paramètres dans le modèle qu'à la condition que ces paramètres apportent un ajout significatif à la vraisemblance. Il faut savoir que l'utilisation de ces critères statistiques représente un champ de recherche actif et controversé en statistiques, dont les tenants et aboutissants ne seront pas développés ici.

## 1.6 Reconstruction de séquences moléculaires ancestrales

Cette section s'attache à présenter les méthodes classiquement utilisées pour reconstruire des séquences ancestrales à partir de modèles de substitution présentés plus haut. Dans la littérature, de nombreuses méthodes existent pour reconstruire des caractères discrets ou continus le long d'un arbre phylogénétique, en utilisant, par exemple, des matrices spécifiant des taux de changements entre états discrets (Pagel and Meade, 2004) ou en supposant que le trait continu évolue selon un processus brownien (Pagel, 1999). La présentation de toutes ces méthodes, quoique passionnantes, dépasseraient le cadre de cette thèse qui s'attache tout particulièrement à la reconstruction ancestrale de séquences moléculaires.

### 1.6.1 Reconstruction de séquences ancestrales par Parcimonie

Le but du maximum de parcimonie est d'identifier les états ancestraux à chaque noeud de l'arbre qui minimisent le nombre de changements hypothétiques entre ces états afin d'expliquer la distribution des états actuels aux feuilles. Fitch (1971) fut le premier à proposer une méthode de reconstruction de séquences ancestrales basée sur le principe de la parcimonie, sachant que cette méthode peut s'appliquer aussi bien aux données nucléiques, protéiques, qu'à n'importe quel type de caractères discrets (morphologiques, etc). Chaque changement d'état est pénalisé uniformément, de telle sorte qu'un changement de l'état $A$ vers l'état $T$ a le même poids qu'un changement de l'état $A$ vers l'état $G$, si l'on prend l'exemple nucléique.

Si l'on se concentre sur la reconstruction ancestrale d'un site particulier d'un alignement, l'algorithme fonctionne en attribuant à chaque noeud de l'arbre un ensemble d'états de caractère qui sont compatibles avec le nombre minimum de changement. L'algorithme fonctionne pour double récursivité, en partant tout d'abord des feuilles de l'arbre et en remontant jusqu'à la racine. Cette traversée de l'arbre est appelée récursion *post-order*. A chaque étape, le *coût* ou nombre de changements $c$ est calculé et l'ensemble des états possibles $S$ déterminé. Pour les feuilles, considérées comme des noeuds de l'arbre, $c = 0$ et l'attribution des états est aisée puisqu'elle correspond tout simplement à l'état observé actuellement. Ensuite, en reprenant les notations de la figure 1.6, pour un noeud interne donné, par exemple $N_1$ dont les deux descendants $N_2$ et $N_3$ ont été précédemment visités, l'attribution des états à ce noeud correspond à l'intersection des deux ensembles d'états des noeuds descendants, si cette intersection n'est pas vide. Dans le cas contraire, c'est l'union des deux ensembles qui est attribué (1). Dans le cas où le nouvel ensemble est une union, un changement est ajouté au nombre de changements déjà pris en compte dans les parties inférieures de l'arbre.

La deuxième récursivité, dite *pre-order*, part de la racine de l'arbre et descend jusqu'aux feuilles. Elle permet l'attribution des états ancestraux finaux aux noeuds internes. Dans le cas

**Algorithme 1** Pré-récursion de l'algorithme de Fitch

**si** $S_{N_2} \cap S_{N_3} \neq \emptyset$ **alors**
  $S_{N_1} \leftarrow S_{N_2} \cap S_{N_3}$
  $c_{N_1} \leftarrow c_{N_2} + c_{N_3}$
**sinon**
  $S_{N_1} \leftarrow S_{N_2} \cup S_{N_3}$
  $c_{N_1} \leftarrow c_{N_2} + c_{N_3} + 1$
**fin si**

où l'ensemble des états possibles à la racine déterminé par la récursion *post-order* est de dimension 1, l'état ancestral final à la racine est égal à l'état présent dans cet ensemble. Dans le cas contraire, plusieurs reconstructions également parcimonieuses existent et l'état ancestral se choisit au hasard dans cet ensemble. Swofford and Maddison (1987) ont proposé deux méthodes d'assignation des états dans le cas d'ambiguïtés : la méthode ACCTRAN pour *acceleration transformation* qui fait l'hypothèse que les changements de caractères ont eu lieu le plus tôt possible dans l'arbre – ce qui a alors tendance à favoriser les réversions au détriment des convergences – et la méthode DELTRAN pour *delayed transformation* qui fait l'hypothèse que les changements ont été tardifs – ce qui favorise alors les convergences.

Au lieu de considérer que tous les changements d'états sont équiprobables comme dans l'algorithme de Fitch (1971), il est possible d'avoir recours à une matrice de coûts arbitraires spécifiant les coûts des différents types de changement d'états. Cela permet de réaliser une reconstruction *pondérée* par parcimonie, plus proche de la réalité biologique. Sankoff (1975) a proposé un tel algorithme, qui est une généralisation de l'algorithme de Fitch (1971) et qui calcule le coût minimum d'un site et énumère les reconstructions qui produisent ce minimum sachant la matrice des coûts. De la même façon que dans l'algorithme de Fitch, l'arbre est parcouru par une récursion *post-order* suivie d'une récursion *pre-order*. Néanmoins, il est difficile de connaitre les poids optimaux à utiliser pour un jeu de séquences donné, ce qui limite l'emploi d'un tel algorithme en pratique.

### 1.6.2 Reconstruction de séquences ancestrales par le biais de modèles probabilistes

#### 1.6.2.1 Reconstruction par Maximum de Vraisemblance

Deux types de reconstruction de séquences ancestrales par Maximum de Vraisemblance sont possibles, à savoir la *reconstruction marginale* ou la *reconstruction jointe* des caractères ancestraux. La reconstruction marginale permet de déterminer pour un site donné à un noeud donné l'état ancestral le plus probable. En revanche, la reconstruction jointe associe à l'ensemble des

$n$ noeuds de l'arbre l'ensemble le plus probable de $n$ états ancestraux.

- **Reconstruction marginale**. Considérons l'arbre de la figure 1.6. Nous nous intéressons à déterminer quel est l'acide aminé ancestral le plus probable au noeud $V_2$ pour le site $C^{(i)}$.

  Pour ce noeud, la probabilité postérieure de l'acide aminé $a$

  $$\mathbb{P}(n_2 = a|C^{(i)}, \theta) = \frac{\mathbb{P}(n_2 = a, C^{(i)}|\theta)}{\mathbb{P}(C^{(i)}|\theta)}$$

  avec $\mathbb{P}(C^{(i)}|\theta)$ la vraisemblance du site et $\mathbb{P}(n_2 = a, C^{(i)}|\theta)$ la probabilité jointe d'avoir l'état $a$ au noeud $N_2$ et les états observés aux feuilles. Cette probabilité peut s'exprimer de cette façon :

  $$\mathbb{P}(n_2 = a, C^{(i)}|\theta) = \mathbb{P}(n_2 = a|\theta) \times \mathbb{P}(C^{(i)}|n_2 = a, \theta)$$

  avec $\mathbb{P}(n_2 = a|\theta)$ la probabilité d'observer l'état $a$ au noeud $N_2$. Si l'on suppose que l'on est à l'état stationnaire, on a : $\mathbb{P}(n_2 = a|\theta) = \pi_a$. $\mathbb{P}(C^{(i)}|n_2 = a, \theta)$ représente la vraisemblance conditionnelle d'avoir les états $C^{(i)}$ aux feuilles sachant l'état $a$ au noeud $N_2$ et les paramètres du modèle. Ces vraisemblances conditionnelles sont calculées à l'aide de l'algorithme d'élagage (*pruning algorithm*) de Felseinstein lors du calcul de la vraisemblance de l'arbre. Bien évidemment, tous ces calculs se font avec les paramètres du modèle et les longueurs de branches préalablement (généralement) estimés au Maximum de Vraisemblance.

- **Reconstruction jointe**. La reconstruction jointe considère tous les noeuds ancestraux en même temps et l'ensemble des états ancestraux attribués aux noeuds pour un site donné ayant la probabilité postérieure maximale est considéré comme la meilleure reconstruction. Ainsi, soit $N = \{a_1, a_2, a_3\}$ un ensemble d'états ancestraux correspondant aux noeuds $N_1$, $N_2$ et $N_3$. Pour un site donné, on a la probabilité postérieure de $N$ qui est égale à :

  $$\mathbb{P}(N|C^{(i)}, \theta) = \mathbb{P}(n_1 = a_1, n_2 = a_2, n_3 = a_3|C^{(i)}, \theta) = \frac{\mathbb{P}(a_1, a_2, a_3, C^{(i)}|\theta)}{\mathbb{P}(C^{(i)}|\theta)}$$

  avec $\mathbb{P}(a_1, a_2, a_3, C^{(i)}|\theta)$ qui est égal à :

  $$\mathbb{P}(a_1, a_2, a_3, C^{(i)}|\theta) = \pi_{a_1} \times \mathbb{P}(a_2|a_1, b_5) \times \mathbb{P}(s_1|a_2, b_1) \times \mathbb{P}(s_2|a_2, b_2) \times \mathbb{P}(a_3|a_1, b_6) \times \mathbb{P}(s_3|a_3, b_3) \times \mathbb{P}(s_4|a_3, b_4)$$

Les séquences ancestrales inférées par reconstruction marginale ou jointe sont souvent similaires. Cependant, il peut arriver que les deux approches proposent des scénarios ancestraux différents, notamment lorsque la divergence moyenne entre les séquences augmente. Il devient alors intéressant de comparer les résultats au niveau des sites variables, surtout si ces séquences sont par la suite clonées et exprimées *in vivo* en laboratoire afin d'en étudier les propriétés biochimiques. Il est fort probable que lorsque les deux méthodes proposent des états différents pour un site à un noeud donné, la reconstruction soit, de fait, incertaine. Cela se reflète alors au niveau des probabilités postérieures de chaque état qui indiquent que plusieurs états ont des probabilités non négligeables. Si les deux approches proposent des scénarios différents pour des reconstructions incertaines, il faut être très prudent vis à vis des conclusions biologiques tirées et, éventuellement lorsque c'est possible, ressusciter les deux protéines en laboratoire afin de vérifier que les incongruences entre les deux méthodes n'ont pas d'influence sur les propriétés biochimiques des protéines ancestrales.

Selon les cas, le calcul des séquences ancestrales par reconstruction jointe peut s'avérer difficile d'un point de vue du temps de calcul requis. Dès que le nombre de séquences dans l'alignement augmente, le nombre de noeuds internes augmente (linéairement) et la combinatoire des scénarios à tester devient rapidement rédhibitoire. C'est encore plus le cas si une distribution $\Gamma$ est utilisée pour modéliser la variation de taux entre sites. Dans ce cas, alors que la reconstruction marginale requiert un temps de calcul qui est linéaire avec le nombre de séquence (avec ou sans loi $\Gamma$ d'ailleurs), la reconstruction jointe est exponentielle avec le nombre de séquences (Pupko et al., 2007). Ainsi, Pupko et al. (2000, 2002) ont développé un algorithme efficace permettant de calculer de manière efficace et exacte la reconstruction jointe de maximum de vraisemblance en considérant une distribution $\Gamma$. En pratique, cette reconstruction jointe peut s'effectuer avec plusieurs dizaines de séquences, ce qui rend accessible son utilisation, notamment dans le but de comparer les reconstructions jointes et marginales.

Plusieurs études ont abordé la comparaison de la reconstruction des séquences ancestrales par MP et ML (Yang et al., 1995; Zhang and Nei, 1997). De manière générale, il a été observé que la reconstruction par parcimonie était bien moins précise que celle au ML. Le fait pour le MP de ne pas prendre en compte les biais de taux de substitutions existant entre bases ou acides aminés dans les séquences biologiques ainsi que de ne pas considérer les variations de taux dans le temps modélisées par les longueurs de branches sont les arguments principaux avancés par ces auteurs. Cependant, lorsque le niveau de divergence entre séquences est faible, l'approche MP montre des performances similaires au ML. Enfin, l'autre désavantage de la parcimonie est de ne pas pouvoir produire d'évaluation de la précision de la reconstruction comme peut le faire le ML à l'aide des probabilités postérieures calculées. Cependant, l'approche MP peut être plus intéressante que l'approche ML dans le cas de reconstruction de caractères discrets

ancestraux comme les caractères morphologiques. En effet, il n'y a souvent pas assez de données pour estimer efficacement les paramètres d'un modèle de changement des états morphologiques, rendant l'approche ML très sensible aux biais stochastiques d'estimation des paramètres et donc d'estimation des états ancestraux.

En sus des reconstructions marginale et jointe, d'autres approches ont été développées pour reconstruire les séquences ancestrales. Par exemple, Schluter (1995) a développé une méthode ML qui optimise conjointement un modèle de substitution spécifique d'un site donné par le biais de la reconstruction des états ancestraux de ce site. La reconstruction maximisant la vraisemblance de ce site conditionnellement aux paramètres de substitution optimisés est considérée comme la meilleure reconstruction. Cette méthode n'est jamais devenue populaire dans la littérature, probablement dû au fait qu'elle requiert l'estimation de nombreux modèles de substitution (un modèle par site) à partir d'une quantité très limitée d'information phylogénétique entrainant potentiellement de nombreux problèmes de reconstruction des états ancestraux. Actuellement, la reconstruction marginale par ML (Yang et al., 1995) est l'approche la plus fréquemment utilisée dans les études abordant la reconstruction et la résurrection de protéines ancestrales (voir plus bas). PAML (Yang, 2007), MEGA (Tamura et al., 2011), DAMBE (Xia, 2013) ou encore bppAncestor (Dutheil and Boussau, 2008) (dépendant des librairies Bio++ (Guéguen et al., 2013)) sont les logiciels ou programmes les plus classiquement utilisés dans la littérature pour effectuer l'inférence de séquences ancestrales.

#### 1.6.2.2 Reconstruction par méthodes bayesiennes

L'approche par Maximum de Vraisemblance requiert l'estimation au Maximum de Vraisemblance de la valeur des paramètres du modèle de substitution et des longueurs de branches. Ceci peut devenir, en théorie, problématique lorsque de petits jeu de données sont analysés. Le faible nombre de sites pouvant entraîner une diminution de l'efficacité de l'optimisation des paramètres, ceci peut se réperctuer sur le calcul des états ancestraux les plus probables. Afin de prendre en compte les erreurs d'échantillonnage des valeurs de paramètres dans le calcul des séquences ancestrales, des approches bayesiennes ont été proposées (Huelsenbeck et al., 2001; Pagel et al., 2004). Ces méthodes utilisent des priors sur les paramètres évolutifs afin d'intégrer les incertitudes sur les valeurs de ces paramètres en échantillonnant dans leur distribution postérieures à l'aide d'algorithme MCMC.

La reconstruction de séquences ancestrales par Maximum de Vraisemblance considère qu'il n'y a pas d'incertitude sur l'arbre phylogénétique, qui est considéré comme vrai. Dans la réalité, cet arbre est souvent irrésolu, notamment à cause de la faible quantité d'information génétique présente dans l'alignement du gène étudié. L'avantage de l'approche est de pouvoir prendre en

compte cette incertitude en reconstruisant les séquences ancestrales en intégrant sur l'espace des topologies. Ainsi, Huelsenbeck et al. (2001) and Hanson-Smith et al. (2010) ont testé l'influence de la prise en compte de cette incertude. Alors que Huelsenbeck et al. (2001) ont montré que les probabilités postérieures pouvaient être affectées par l'incertitude sur l'arbre, Hanson-Smith et al. (2010) ont montré sur données simulées et réelles que cela avait un impact très mineur sur la détermination de l'état de maximum de probabilité postérieure.

Dans la réalité, il semble à première vue que les reconstructions par approches de Maximum de Vraisemblance et approches Bayesiennes ne diffèrent que légèrement. Seulement, il faut faire très attention sur l'effet que cela peut avoir sur les conclusions biologiques. Hanson-Smith et al. (2010) ont par exemple conclu qu'il ne servait à rien de prendre en compte l'incertitude sur la topologie. Non seulement cette conclusion peut être criticable étant donnés les simulations peu réalistes et les arbres de gènes globalement bien soutenus utilisés, mais selon le contexte biologique et la protéine étudiée, quelques changements d'acides aminés seulement peuvent modifier les propriétés de stabilité ou de fonctionnalité des ancêtres ressuscités.

### 1.6.3 Biais de reconstruction des séquences ancestrales par parcimonie et Maximum de Vraisemblance

Les méthodes de parcimonie et de Maximum de Vraisemblance souffrent d'un biais systématique d'estimation des états ancestraux, notamment lorsque les états de maximum de parcimonie ou de probabilité postérieure maximale sont considérés. Le biais réside entièrement dans le fait d'utiliser la séquence de probabilité maximale et d'ignorer les solutions sous-optimales (Yang, 2006). Ce biais résulte en l'enrichissement en acides aminés fréquents dans les séquences ancestrales au fur et à mesure que l'on s'éloigne des feuilles. Pour un site donné, les approches MP et ML vont avoir tendance à attribuer de manière biaisée aux noeuds ancestraux les caractères les plus souvent rencontrés au niveau des sites, c'est à dire les caractères ayant une forte fréquence observée (spécifique du site), tout en excluant les caractères moins fréquents. La plupart du temps, les sites contraints fonctionnellement sont conservés à travers le temps, de telle sorte qu'ils sont occupés par des états dit "favorables" pour la protéine, qui sont donc à forte fréquence, alors que les états moins favorables vont être observés à faible fréquence. La conséquence du biais mentionné ci-dessus est que les protéines reconstruites vont progressivement s'appauvrir en acides aminés moins favorables lorsque l'on s'éloigne des feuilles. Ainsi, les protéines ancestrales de ces protéines comme la thermostabilité ou les constantes cinétiques risquent d'être non seulement mal estimées, mais surtout systématiquement surestimées. Williams et al. (2006) ont mis en évidence ce phénomène. En simulant des séquences protéiques selon un modèle thermodynamique de mutation-sélection contraignant les séquences à conserver une stabi-

lité optimale selon un modèle structural, ils ont pu comparer les stabilités thermodynamiques des protéines reconstruites à celles enregistrées pendant la simulation. Ils ont comparé les approches MP, ML et bayesienne et ont remarqué que les méthodes MP et ML reconstruisaient systématiquement des protéines ayant une stabilité thermodynamique supérieure aux stabilités "vraies". Par contre, l'approche bayesienne n'est pas affectée par ce biais car les résidus moins favorables sont pris en compte dans la distribution postérieure des séquences ancestrales et seront présents avec une fréquence non biaisée. Cependant, les stabilités thermodynamiques obtenues dans cette étude à partir des séquences primaires de protéines sont théoriques et calculées à partir de potentiels statistiques modélisant l'énergie libre de la séquence, en supposant une conservation de la structure au cours du temps. Beaucoup de biochimistes considèrent ces prédictions comme étant irréalistes et non fiables, insistant sur le fait qu'il est extrêmement difficile d'estimer la thermostabilité d'une protéine à partir de sa séquence primaire (Joanne Hobbs, communication personnelle).

À ce biais s'ajoute un autre biais pour la reconstruction en ML. En effet, l'approche ML va avoir tendance à incorporer de manière biaisée dans les séquences ancestrales l'acide aminé ayant la plus forte fréquence dans le modèle. En fonction de cet acide aminé et de la protéine considérée, les propriétés biologiques des séquences ancestrales peuvent être impactées. En revanche, la méthode bayesienne est robuste face à ce biais.

Ces observations ont poussé plusieurs chercheurs à suggérer de ne pas utiliser uniquement la séquence ayant la probabilité maximale en ML lorsque des résurrections de ces protéines sont envisagées afin d'en étudier les propriétés (Williams et al., 2006). Au contraire, il est recommandé de tirer aléatoirement quelques séquences dans la distribution postérieure des séquences ancestrales et de toutes les caractériser expérimentalement afin de vérifier que les caractéristiques biologiques sont conservées parmi ces variants. Malgré cela, l'approche ML reste l'approche produisant les séquences ancestrales les plus précises, c'est à dire effectuant le moins d'erreur de reconstruction. De futures expériences sont nécessaires afin de développer des approches ML robustes à ce type de biais. Une des possibilités est d'utiliser des modèles non-homogènes dans le temps qui optimisent des fréquences d'équilibres variant le long de l'arbre. Si l'acide aminé majoritaire dans le modèle change d'une branche à l'autre ou d'une région à l'autre de l'arbre, le deuxième biais mentionné ci-dessus ne devrait plus se propager le long de l'arbre lors de la reconstruction. Néanmoins, bien que cette approche soit intéressante, elle ne reste justifiable qu'à partir du moment où les compositions globales des séquences analysées varient, ce qui n'est pas toujours le cas.

### 1.6.4 Reconstruction de séquences ancestrales et applications

La reconstruction de séquences ancestrales s'avère être une approche très intéressante pour la production de nouvelles protéines à visée industrielle. La conception, par la suite appelée *design*, de protéines peut permettre la création de bibliothèques de séquences protéiques possédant des propriétés fonctionnelles ou structurales optimisées par rapport à ce qui peut exister dans la nature (Cole and Gaucher, 2011). Plusieurs techniques existent pour créer à partir de séquences actuelles une vaste quantité de protéines ayant des caractéristiques différentes. Par exemple, la construction de séquences consensus à partir de séquences actuelles ou l'incorporation d'acides aminés consensus dans des protéines actuelles par mutagénèse dirigée a par le passé été utilisé avec succès (Cole and Gaucher, 2011), comme dans le cas du design de protéines ayant de meilleures stabilités thermodynamiques. Lehmann et al. (2002) ont calculé des séquences consensus de phytases à partir de séquences de champignons. La phytase permet l'hydrolyse d'acide phytique, molécule représentant la forme majeure de stockage de phosphore dans les graines de plantes et qui n'est pas digérée par les animaux non-ruminants. Les protéines consensus obtenues ont une stabilité thermodynamique bien supérieure aux séquences actuelles, ce qui peut avoir un avantage important pour une utilisation à large échelle de ces protéines en industrie. Au passage, ces auteurs ont exploité, probablement sans le savoir, le biais de reconstruction mentionné dans la section 1.6.3, qui fait que les acides aminés "favorables" se retrouve majoritairement dans la séquence consensus. D'autres techniques de design de protéines existent, comme le DNA shuffling, consistant à fragmenter des séquences homologues d'ADN codant puis de les recombiner aléatoirement à l'aide de PCR (Ness et al., 2002).

Une des problématiques rencontrée par le design de protéines est l'énorme combinatoire des arrangements d'acides aminés possibles et la sélection des protéines d'intérêt (Denault et al., 2007). Cet espace protéique est bien plus vaste que l'espace des protéines effectivement fonctionnelles. Les pressions sélectives imposant aux protéines un repliement, une fonctionnalité et une stabilité efficaces expliquent cette réduction de l'espace des possibles. Étant donné que les séquences actuelles ont été par le passé sujettes à ces pressions de sélection, l'exploration de l'espace des séquences ancestrales garantit plus facilement la construction de bibliothèques à la fois plus restreintes et contenant des protéines fonctionnelles. Par exemple, Bershtein et al. (2008) ont utilisé la reconstruction de séquences ancestrales de TEM-1 $\beta$-lactamase (enzyme produite par des bactéries Gram négatives leur conférant une résistance à la pénicilline) pour montrer que les ancêtres des séquences actuelles étaient plus stables et plus à même d'acquérir de nouvelles fonctions. Cela suggère l'idée de les utiliser comme séquences 'parentes' lors d'études d'évolution expérimentale en laboratoire qui ont pour but de produire une grande quantité de variants géniques *in vitro*, en imposant de forts taux mutationnels.

L'utilisation de séquences ancestrales à visée industrielle n'en est encore qu'à ses prémices. Néanmoins, cette approche a déjà fourni des résultats très prometteurs et peut permettre d'envisager la synthèse de protéines plus stables que les protéines actuelles dans des environnements extrêmes tels que des environnements à haute température, acidité ou salinité.

## 1.7 Description brève des articles présentés dans cette thèse.

Ci-dessous sont brièvement décrits les différents articles que je présente dans les trois prochaines grandes parties (parties 2, 3 et 4). Ces articles sont soit déjà publiés au moment de la rédaction de ce manuscrit ou à l'état de manuscrit soumis ou non-soumis. Les numérotations ci-dessous correspondent aux numérotations des chapitres et sections présentant ces articles dans la suite du manuscrit.

2. De nouveaux modèles de substitution protéiques hétérogènes en ML.

    2.1. A Branch-Heterogeneous Model of Protein Evolution for Efficient Inference of Ancestral Sequences.
    Cet article présente le développement d'un nouveau modèle de substitution hétérogène en composition, applicable sur des données protéiques en Maximum de Vraisemblance. Ce modèle, appelé COaLA, permet de modéliser la variation de compositions globales en acides aminés entre lignées et utilise une analyse de correspondences pour réduire l'espace des paramètres à optimiser.

    2.2. Efficient modeling of protein site-heterogeneities with empirical mixtures of profiles.
    Ce manuscrit détaille un nouveau jeu de modèles empiriques hétérogènes en sites. Ces modèles, nommés ECG, sont des modèles de mélange de profiles et permettent de prendre en compte la variation du processus évolutif entre sites.

3. Renaissance *in silico* et évolution précoce du monde microbien.

    3.1. Adaptation to Environmental Temperature Is a Major Determinant of Molecular Evolutionary Rates in Archaea.
    Dans cet article, les températures ancestrales de vie sont reconstruites le long de l'arbre phylogénétique des Archées. Un résultat majeur de ce papier est la mise en évidence du rôle de l'adaptation à la température au cours du temps dans la variation des taux d'évolution entre espèces.

    3.2. The molecular signal for the adaptation to cold temperature during early life on Earth.
    Plusieurs résultats précédents ont suggéré que le dernier ancêtre commun à toutes les espèces actuelles vivait à basse température, contrairement à deux de ses descendants proches, à savoir les ancêtres des Bactéries et des Archées. Cet article présente la nature du signal phylogénétique capté par les modèles d'évolution pour proposer un tel scénario non-parcimonieux d'adaptation à la température au niveau des lignées les plus profondes de l'arbre de la vie.

3.3. Ribosomal proteins as next generation standard for prokaryotic systematics.
Ce manuscrit présente l'intérêt de l'utilisation des protéines ribosomales pour la reconstruction phylogénétique des espèces procaryotes. Ceci est illustré par la reconstruction phylogénétique des Protéobactéries. Le manuscrit présente aussi l'utilisation de modèles hétérogènes en compositions et de détection de transferts horizontaux de gènes afin de raciner l'arbre des Protéobactéries.

4. La résurrection de protéines ancestrales.

4.1. Biologically motivated models strongly improve the functionality of resurrected proteins.
Ce manuscrit présente un nouveau protocole de reconstruction de séquences protéiques ancestrales dans le but de les ressusciter en laboratoire. Il présente l'importance d'utiliser des modèles de substitutions hétérogènes et des arbres de gènes réconciliés lors du calcul des séquences ancestrales les plus probables et valide ces résultats *in silico* par la résurrection d'une enzyme impliquée dans la synthèse de la Leucine.

4.2. Resurrection of halophilic proteins provides insights into the evolution of protein structure and function.
Ce manuscrit décrit la reconstruction et la résurrection du gène de la malate déhydrogénase au sein des archées halophiles afin de comprendre le lien entre la structure et la fonction des protéines lors de l'adaptation aux environnements halophiles.

En annexe, deux autres articles publiés sont également présentés. Le premier décrit la nouvelle version des librairies Bio++ dans lesquelles les nouveaux modèles présentés durant cette thèse ont été développés. Le second résulte d'un stage effectué lors de ma dernière année de Master dans le laboratoire de Ziheng Yang et traite de l'estimation des temps de divergence le long de l'arbre des Foraminifères benthiques.

# 2

# De nouveaux modèles de substitution protéiques hétérogènes en ML.

## 2.1 Le modèle COaLA : modélisation efficace de la variation de composition globale dans le temps.

### 2.1.1 Introduction

Comme expliqué précedemment en introduction, les modèles hétérogènes en temps permettent de modéliser les variations de la composition globale des séquences dans le temps. Galtier and Gouy (1998) puis Boussau and Gouy (2006) ont implémenté un tel modèle en nucléique en ML. Dans ce modèle, le taux global en bases G+C varie de branche à branche. Modéliser des variations globales de compositions de branche à branche est très probablement irréaliste, mais c'est une hypothèse qui facilite grandement l'implémentation de modèles hétérogènes en temps en ML. Pour des données protéiques, en ML, seul bppML (Dutheil and Boussau, 2008) donnait la possibilité d'utiliser un modèle hétérogène en temps avec des jeux de fréquences d'équilibres spécifiques des branches. Cependant, cela nécessitait l'optimisation de $19 \times b$ paramètres, correspondant à l'ensemble des fréquences d'équilibres le long d'un arbre contenant $b$ branches. Cela

rendait l'utilisation d'un tel modèle impossible, pour des raisons pratiques de temps de calcul et des raisons d'efficacité d'optimisation des paramètres.

Dans l'article qui suit, publié dans le journal *Systematic Biology*, un modèle hétérogène en temps efficace pour le ML est présenté, ainsi que tous ses avantages concernant l'amélioration des inférences phylogénétiques.

### 2.1.2 Manuscrit

*Syst. Biol.* 62(4):523–538, 2013

DOI:10.1093/sysbio/syt016
Advance Access publication March 7, 2013

# A Branch-Heterogeneous Model of Protein Evolution for Efficient Inference of Ancestral Sequences

M. Groussin[1,*], B. Boussau[1,2], and M. Gouy[1]

[1]*Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, Université Lyon 1, CNRS, UMR5558, Villeurbanne, France; and* [2]*Department of Integrative Biology, University of California, Berkeley, CA, USA*

[*]*Correspondence to be sent to: Laboratoire de Biométrie et Biologie Évolutive, Université Lyon 1, 43 bd du 11 novembre 1918, UMR CNRS 5558, 69622 Villeurbanne cedex, France; E-mail: mathieu.groussin@univ-lyon1.fr.*

*Received 18 June 2012; reviews returned 23 August 2012; accepted 2 March 2013*
*Associate Editor: Lars Jermiin*

*Abstract*.—Most models of nucleotide or amino acid substitution used in phylogenetic studies assume that the evolutionary process has been homogeneous across lineages and that composition of nucleotides or amino acids has remained the same throughout the tree. These oversimplified assumptions are refuted by the observation that compositional variability characterizes extant biological sequences. Branch-heterogeneous models of protein evolution that account for compositional variability have been developed, but are not yet in common use because of the large number of parameters required, leading to high computational costs and potential overparameterization. Here, we present a new branch-nonhomogeneous and nonstationary model of protein evolution that captures more accurately the high complexity of sequence evolution. This model, henceforth called Correspondence and likelihood analysis (COaLA), makes use of a correspondence analysis to reduce the number of parameters to be optimized through maximum likelihood, focusing on most of the compositional variation observed in the data. The model was thoroughly tested on both simulated and biological data sets to show its high performance in terms of data fitting and CPU time. COaLA efficiently estimates ancestral amino acid frequencies and sequences, making it relevant for studies aiming at reconstructing and resurrecting ancestral amino acid sequences. Finally, we applied COaLA on a concatenate of universal amino acid sequences to confirm previous results obtained with a nonhomogeneous Bayesian model regarding the early pattern of adaptation to optimal growth temperature, supporting the mesophilic nature of the Last Universal Common Ancestor. [Ancestral sequence reconstruction; nonhomogeneous model; optimal growth temperature; phylogenomics; phylogeny.]

Many evolutionary studies use genomic sequences to infer a phylogenetic tree depicting the relationships between species. To reconstruct such trees, substitution models that describe the stochastic process of evolution acting on sequences are preferred. The use of complex models of evolution has provided insights into early events of evolution such as the origin of major groups of organisms (Cox et al. 2008; Philippe et al. 2011), the absolute or relative chronological appearance of major clades or important phenotypic characters (Douzery et al. 2004; Delsuc et al. 2006), and ancestral conditions of life (Boussau and Gouy 2012). Over recent years, many authors have proposed to perform ancestral sequence reconstruction to tackle such problems, either at the scale of a single gene alignment (Gaucher et al. 2008; Finnigan et al. 2012) or at the scale of concatenates of genes (Boussau et al. 2008; Groussin and Gouy 2011). To infer the characteristics of ancestral molecules from the analysis of extant genomes, accurate and biologically relevant models of evolution must be utilized.

However, standard models are usually designed with the simplifying assumptions that the evolutionary process was globally stationary, reversible, and homogeneous (Yang 2006; Jermiin et al. 2008; Jayaswal et al. 2011a) (Fig. 1a). It has been shown that homologous sequences can diverge widely in their base or amino acid compositions (Hasegawa and Hashimoto 1993; Galtier and Lobry 1997; Zeldovich et al. 2007). Consequently, the assumption that the composition of nucleotides or amino acids in the sequences has remained unchanged from the root of the tree to its leaves (stationarity hypothesis), and that all branches of a phylogenetic tree share the same relative amino acid substitution rates (homogeneity hypothesis), is not appropriate for compositionally heterogeneous sequences. Compositional heterogeneity across sets of homologous sequences may lead to erroneous reconstructions of phylogenetic trees or ancestral frequencies (Ho and Jermiin 2004; Jermiin et al. 2004; Blanquart and Lartillot 2006, 2008; Boussau and Gouy 2006; Boussau et al. 2008). A natural approach to avoid these erroneous reconstructions is to use a model that represents in a more realistic fashion the evolutionary process.

Several models that relax the homogeneity and stationarity hypotheses have been developed, either in the distance-based framework (Lake 1994; Lockhart et al. 1994; Galtier and Gouy 1995; Tamura and Kumar 2002) or in the likelihood or Bayesian frameworks (Yang and Roberts 1995; Galtier and Gouy 1998; Foster 2004; Jayaswal et al. 2005, 2007, 2011b; Blanquart and Lartillot 2006; Dutheil and Boussau 2008; Zou et al. 2012). In each of these methodological contexts, the branch-heterogeneous models require several substitution matrices to be used for a given phylogenetic tree (Fig. 1b) whereas the branch-homogeneous models only require one such matrix (Fig. 1a). Therefore, more parameters need to be estimated for branch-heterogeneous models than for branch-homogeneous models. The purpose of branch-heterogeneous models is to decrease the bias in the estimation of model parameters, but their drawback may be an increase in variance. This trade-off between bias and variance should be a matter

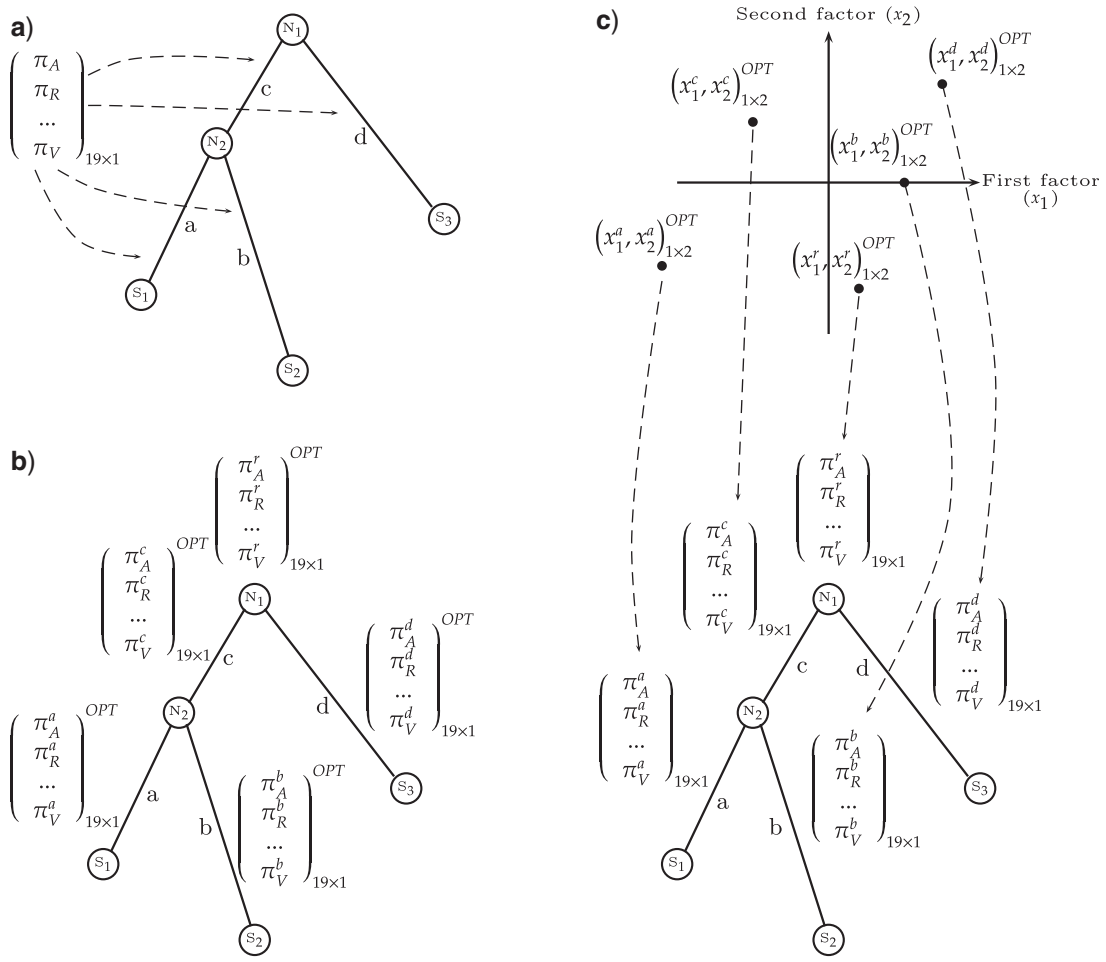



FIGURE 1. The COaLA model substantially decreases the dimension of the space of equilibrium frequency parameters. a) In homogeneous and stationary models, only one vector of amino acid frequencies represents the equilibrium state of sequences and is used for likelihood computation. This vector may be optimized by ML (LG+$F_{opt}$ model) or not (LG or LG+$F_{obs}$ models). b) With a standard nonhomogeneous approach, the homogeneity and stationarity hypotheses are relaxed by assigning independent vectors of 19 equilibrium frequencies per branch to model the variations of overall composition through time. c) With the COaLA model, small dimension vectors of coordinates along the first axes of the COA are optimized per branch. In this example, a two-dimension vector corresponding to the first two axes is associated to each branch and is optimized by ML (OPT). Reversing the COA (dashed arrows), from a vector of coordinates in the low-dimension space, one can compute the corresponding vector of 20 frequencies that is used to compute transition probabilities along the branch.

of concern when employing parameter-rich models (Wertheim et al. 2010). It is necessary to make sure that the parameters that capture the time variability of global compositions increase the fit of the model to the data enough to compensate for the increased number of parameters. Objective criteria such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) can be used to determine the optimal choice for the trade-off between the fit of the model to the data and the number of parameters in the model (Steel 2005). Thus, it was observed that branch-heterogeneous models of sequence evolution may be preferred or rejected over branch-homogeneous models, depending on the choice of parameters or the amount of heterogeneity in the data (Dutheil and Boussau 2008; Groussin and Gouy 2011). Finally, the issue of computational cost has hampered the use of branch-heterogeneous models at a broad scale, especially for proteins, making the development of statistically and computationally efficient branch-heterogeneous models necessary. Note that for convenience the terms "branch-heterogeneous" and "nonhomogeneous" are used interchangeably in the rest of the article, excepted in cases where "nonhomogeneous" is used to describe other types of heterogeneities (e.g., site-specific process-heterogeneity).

Several studies have presented approaches to reduce the number of parameters to be estimated with branch-heterogeneous models. For instance, some methods do not estimate one matrix per branch, but use groups of branches that share substitution matrices (Yang 1998; Foster 2004; Dutheil and Boussau 2008). These groups can be defined *a priori* (Dutheil and Boussau 2008), or estimated during the course of the computation (Jayaswal et al. 2011a; Dutheil et al. 2012). Similarly, Bayesian approaches have been developed that place

breakpoints along the branches of the phylogeny: substitution models are shared by all branches between breakpoints, but change at breakpoints (Blanquart and Lartillot 2006, 2008). Another approach to further reduce the number of parameters has been to share some parameters of the substitution matrices among all branches and have only a subset of them estimated separately for each branch or group of branches. Using such an approach, Galtier and Gouy (1998) were able to propose a branch-heterogeneous model of nucleotide sequence evolution with only one extra parameter per branch of the phylogenetic tree, namely branch-wise equilibrium G+C contents. The resulting model has a good fit to the data because some nucleotide sequences vary extensively in their G+C content. For amino acid sequences, however, it is unclear how variations among homologous sequences could be efficiently summarized by a single or even a small number of variables for any data set.

An efficient model of protein evolution would be useful in studies aimed at protein resurrection. Ancestral sequence reconstruction and resurrection is a powerful approach to characterize ancient molecular properties, to highlight the complex relationship between sequence, structure, and function, or to infer past lifestyle conditions (Harms and Thornton 2010; Boussau and Gouy 2012). The widely applied protocol for ancestral sequence reconstruction starts with the choice of one of the time-reversible Markov models that ModelTest (Posada and Crandall 1998; Posada 2008) considers. Then, this model is used in the PAML package (Yang 2007) to compute the most likely ancestral sequence at each internal node of a maximum-likelihood (ML) tree for the gene under consideration. However, the variation of the substitution process through time or among sites is not accounted for, even when billions of years separate all sequences from their common ancestor (Gaucher et al. 2008; Hobbs et al. 2011). Using a model that can take into account a higher proportion of the complexity of evolutionary processes without excess of variance should help inferring better ancestral sequences.

We do not know of any statistically and computationally efficient branch-heterogeneous substitution model for proteins in the ML framework. Here, we present the correspondence and likelihood analysis (COaLA) model, a new branch-heterogeneous model of amino acid sequence evolution for ML. This model achieves computational efficiency through the same means as Galtier and Gouy (1998), reducing the number of variables that need to be estimated per branch of a phylogenetic tree; it focuses only on a few directions explaining most of the compositional variance observed in the data (Fig. 1c). These variables correspond to linear combinations of observed amino acid frequencies in the data set according to a correspondence analysis (COA) (Greenacre 1984). COA constructs linear combinations of amino acid frequencies ranked by decreasing contribution to the explained variance (these linear combinations are also called axes or factors in the statistical literature; here, we refer to them as axes). Consequently, exploring different values along the first axes amounts to exploring a high proportion of the compositional variability encountered in the data set. In addition, as COA has been previously used to characterize the determinants of compositional heterogeneity among protein sequences (Boussau et al. 2008), estimated branch-wise values along the axes of the COA may be used to directly gain information about the evolution of biological or physical properties affecting compositions over time.

In this article, we describe the COaLA model and how it is applied to the data. The model has been tested on both simulated and biological data sets and we show results focusing on its ability to efficiently fit the data, estimate ancestral frequencies as well as ancestral sequences in comparison with standard homogeneous models. Finally, we apply the model on a previously published data set to confirm the phylogenetic signal explaining the early pattern of adaptation to environmental temperature before the emergence of the three domains of life.

## Materials and Methods

### *Branch-Homogeneous and Branch-Heterogeneous Markovian Substitution Processes*

We consider a tree, $T$, rooted at node $r$, along which amino acid sequences evolve. Sequence evolution proceeds from the root to the leaves of the tree, where sequences are observed. At the root, a vector $\pi_r$ specifies the amino acid frequencies of the (unobserved) ancestral sequence. Along the branches of the tree, we assume that substitutions occur according to a Markov process. In the context of molecular evolution, the kernel of the Markov process is called the substitution matrix and is denoted as $\mathbf{Q}$. If the kernel is time-reversible, then $\mathbf{Q}$ can be decomposed into two matrices, $\rho$ and $\mathbf{\Pi}$, where $\rho = \rho_{yz}$ is a matrix of exchangeabilities (or relative exchange rates) and $\mathbf{\Pi} = \mathrm{diag}(\pi_y)$ is the diagonal matrix of stationary or equilibrium frequencies (Whelan and Goldman 2001), with $y, z = 1, \ldots, 20$ (where 20 is the number of amino acids). The general term of $\mathbf{Q}$ is computed as follows:

$$\mathbf{Q}_{y,z} = \rho_{yz}\pi_z, \text{ with } y \neq z$$

$$\mathbf{Q}_{y,y} = -\sum_{z \neq y} \mathbf{Q}_{y,z},$$

with $\rho_{yz} = \rho_{zy}$ for $y > z$.

The transition probabilities $p_{y \to z}(t)$, defined as the probability of change from state $y$ to state $z$ along an edge of length $t$ time units, are calculated as follows:

$$p_{y \to z}(t) = \left[ e^{\mathbf{Q}t} \right]_{yz}.$$

Common models of sequence evolution assume a constant substitution rate matrix over the tree (Jermiin et al. 2008). Such models are said to be globally homogeneous. In addition, it is often assumed that the

evolutionary process is at equilibrium, in which case the process is also said to be stationary with $\pi_r = \pi$. Reversibility implies that the flux from one amino acid $y$ to another $z$ is equal to the flux from $z$ to $y$:

$$\pi_y p_{y\to z}(t) = \pi_z p_{z\to y}(t).$$

These assumptions have two major consequences: (i) such models (i.e., the commonly used models of sequence evolution) cannot infer a direction of evolution, so the root can be placed anywhere on the tree without affecting the likelihood value (Felsenstein 1981; Yang 2006) and (ii) as previously noted (Galtier and Gouy 1998; Boussau and Gouy 2006), these models assume that all sequences in a tree share similar base or amino acid frequencies.

As illustrated in Jermiin et al. (2008), the evolutionary process can be defined as one of the six (out of eight) possible permutations of homogeneous/nonhomogeneous condition, reversible/nonreversible condition, and stationary/nonstationary condition. These conditions may be applied globally (e.g., to every branch in the tree) or locally (e.g., to a particular branch of the tree). The model presented here is designed to work on amino acid data and to relax the assumptions of global homogeneity, reversibility, and stationarity; in other words, it allows different lineages to diverge toward different amino acid compositions, starting from another set of amino acid frequencies at the root ($\pi_r$). Therefore, the model is nonreversible, and the position of the root affects the likelihood value. COaLA is inspired from the N2 model initially proposed by (Yang and Roberts 1995), designed for DNA, in which a single exchangeability matrix is shared by all branches of $T$, and a distinct vector of equilibrium nucleotide frequencies is associated with each branch of $T$. The model also uses the vector $\pi_r$ of amino acid frequencies at the root.

### *Mathematical Model*

COA is a standard multivariate statistical technique that decomposes the $\chi^2$ statistic associated with a contingency table into orthogonal factors that represent most of the variance (Thioulouse et al. 1997). Here, the contingency table is the matrix of observed amino acid frequencies in protein sequences. In essence, COA summarizes the original data variability using a reduced number $k < 20$ of variables (the factors or axes), which are linear combinations of the 20 original frequencies (see Appendix). Thus, COA reveals the principal axes of a high-dimensional space, enabling at the end the projection of amino acid frequencies into a subspace of lower dimension. In that sense, COA is similar to principal component analysis (PCA). However, PCA uses the Euclidean distance between vectors of frequencies, whereas COA uses the $\chi^2$ distance, which makes COA equally sensitive to deviations in rare amino acids as it is to deviations in frequent amino acids.

The compositional variation among all compared protein sequences is thus summarized in a subspace capturing most of this variation. This subspace allows us to reduce the dimension of the above-mentioned branch-heterogeneous model of protein evolution along a tree by working in the subspace of $k$ principal axes instead of the complete space of 20 parameters. The dimension of the evolutionary model with branch-specific equilibrium frequencies is thus reduced from 19 free parameters per branch to $k$ per branch. From a set of coordinates on a chosen number $k$ of principal axes, it is possible to reverse the COA in order to compute a 20-dimensional vector of amino acid frequencies for which the COA would give these coordinates as factor values (see Appendix). The reduced evolutionary model works by optimizing $k$ coordinates on each branch of $T$, which are transformed into branch-specific vectors of equilibrium amino acid frequencies, which in turn define branch-specific substitution matrices. To illustrate this, consider a rooted phylogenetic tree of 30 species, containing 58 branches and imagine a full branch-heterogeneous LG+$F_{opt}$ model, where 19 free frequencies are optimized per branch and on the root, with a common LG exchangeability matrix (Le and Gascuel 2008) for all branches. It can be compared with a branch-heterogeneous COaLA model where only $k$ free parameters ($k \in [1:19]$) representing axis positions are estimated per branch and on the root. In the first case, the number of parameters ($m$) involved in the model is $m = 19 \times 58 + 19 = 1121$, whereas in the second case, the number of parameters is $m = k \times 58 + k$. As most of the COA performed on real alignments show that a large majority of the variance is explained by the first two or three axes, the improvement in terms of number of parameters can be huge. Thus, if $k = 2$, $m = 118$, a number of parameters 10-fold smaller than with the full approach.

### *Model Availability*

The COaLA model is implemented in the Bio++ libraries (Dutheil et al. 2006), which are a set of freely available C++ libraries dedicated, among other things, to evolutionary biology. The model can be employed with the BppML program, available in the bppSuite series of programs (Dutheil and Boussau 2008). BppML is a general program to optimize a large set of homogeneous/stationary or nonhomogeneous/nonstationary models in the ML framework for several types of data sets (e.g., DNA, codons, and proteins). Information on the model and on how to download and install the libraries can be found at http://pbil.univ-lyon1.fr/software/COaLA/.

### *Models Used in This Study*

In the following, homogeneous and stationary, homogeneous and nonstationary, and nonhomogeneous and nonstationary approaches will be referred by H–S, H–NS and NH–NS approaches, respectively. For

all phylogenetic experiments, the LG exchangeability matrix (Le and Gascuel 2008) is used, but every empirical exchangeability matrix may be employed (e.g., JTT [Jones et al. 1992]; WAG [Whelan and Goldman 2001]). When the vector of equilibrium frequencies specific to the LG model is employed, we will refer to the model as LG. If the vector of equilibrium frequencies is fixed to the observed frequencies computed from the alignment under study (the so-called "+F" model [Adachi and Hasegawa 1996]), the model is referred as LG+$F_{obs}$. When stationary frequencies are optimized by ML, LG+$F_{opt}$ is used. COaLA can also be used as an H–S model. If so, LG+COaLA[$k$] means that the equilibrium frequencies of the single substitution matrix in use by all branches are optimized through $k$ axis positions. With an H-NS approach, a second and independent set of axis positions is optimized on the root. With an NH approach, LG+COaLA[$k$] means that $k$ independent axis positions per branch and on the root are optimized. In this study, the number of axis positions $k$ is set *a priori* and is equal for all branches of the tree. This number is not optimized during the run of the program. Rather, the method is run with all integer values between 1 and $k$, and the optimal number of axes is then determined according to model selection statistical criteria (AIC or BIC, see below). Note that the method could be generalized so that $k$ is optimized to obtain variable numbers of axis positions per branch.

## Simulations

### *Sequence Simulations*

All simulations of amino acid sequences with nonhomogeneous models were performed with BppSeqGen, from the bppSuite series of programs (Dutheil and Boussau 2008).

To simulate these nonhomogeneous amino acid sequences, we considered the 5000 trees used by Guindon and Gascuel (2003) to test the performance of PhyML and which are available at http://www.atgc-montpellier.fr/phyml/datasets.php. These trees contain 40 species. We randomly removed 20 of these 40 species for each of the 5000 trees. Branch lengths were increased to allow different parts of the tree to have sufficient time to diverge in terms of compositions. Thus, the height of the tree, defined as the maximum distance between a leaf and the root, was set to a minimum of 0.8 substitutions/site and all other branches were scaled up accordingly. The resulting branch lengths are still realistic since the overall mean is 0.13 substitutions/site/edge and the overall median is 0.08 substitutions/site/edge, showing that many small branches remain in the trees. We simulated alignments of 5000 amino acids, with rate heterogeneity across sites modeled by a discretized $\Gamma$ distribution with four rate categories (Yang 1994). To specify the nonhomogeneity and nonstationarity, we assigned different independent sets of amino acid equilibrium frequencies to different parts of the tree as well as one for the root; these sets of frequencies were drawn from a Dirichlet distribution. To do so, we determined the means and standard deviations of each amino acid frequency from a protein sequence alignment containing 3336 sites from 115 species spanning the tree of life (Boussau et al. 2008). These means and standard deviations were used to define the marginal densities employed to randomly draw the sets of equilibrium frequencies from the Dirichlet distribution. For each amino acid, we multiplied the observed standard deviations by 3 to increase the nonhomogeneity of simulated sequences in terms of composition. Only two or four different parts of the tree are specified to have different equilibrium compositions (see below), all branches belonging to one of these parts being compositionally homogeneous. This procedure was adopted in order to generate alignments with sizeable levels of compositional heterogeneity. In addition, we randomly drew a set of frequencies that was assigned to the root. We then randomly chose an integer number $w$ (1 or 2). If $w=1$, independent sets of frequencies were assigned on the first two branches around the root. If $w=2$ and if the root has four descendant nodes, the first six branches were assigned different equilibrium compositions. Finally, all branches below one of the nodes of the $w$-th generation were assigned the set of frequencies of the preceding branch leading to that given node.

For each of these 5000 simulated alignments, we computed all pairwise Bowker tests (Bowker 1948) to assess the global heterogeneity of the alignment. The Bowker test relies on a pairwise comparison and on a test of symmetry between two aligned sequences (Ababneh et al. 2006). If the test statistic from the Bowker test is significant, then it is unlikely that the pair of diverging sequences being considered have evolved under the same process. As Dutheil and Boussau (2008) proposed, we defined the global heterogeneity of the alignment as the number of tests that are statistically significant at the 5% level that we corrected with a Holm–Bonferroni correction (Holm 1979) for multiple test comparisons.

Many among the 5000 alignments were moderately heterogeneous according to the Bowker test (half of all the alignments had less than 37% significant pairwise tests). To globally assess the ability of NH–NS COaLA to estimate ancestral frequencies and branch lengths regardless of the data heterogeneity, the 1000 (out of 5000) first trees were selected and their corresponding alignments were analyzed. Moreover, to compare the fit to the data between COaLA and H–S approaches, we retrieved the alignments having the top 5% highest heterogeneity among the 5000 alignments. The mean heterogeneity of the resulting 272 alignments was in accordance with what is observed on empirical data (about 64% of the tests were significant, which is comparable with the heterogeneity of the biological data sets used in this study [see below] and many other concatenated protein data sets [data not shown]).

### *Assessing the Performance of COaLA on Simulated Sequences*

To globally assess the performance of NH–NS COaLA, we first focused on (i) its ability to estimate ancestral frequencies and (ii) to fit data.

We evaluated the capacity of different models to reconstruct sequence evolution from simulated alignments by two means. First, we investigated the accuracy of the reconstructed amino acid frequencies at the root. Second, we evaluated the capacity of the models to reproduce the composition of simulated alignments, in a manner akin to parametric bootstrapping or posterior predictive simulations (Huelsenbeck et al. 2001; Bollback 2002; Lartillot and Philippe 2004). We ran each model on each simulated alignment, and recorded the estimated parameters. Then, we used these parameters to simulate new alignments using BppSeqGen. Finally, we compared these newly simulated alignments with the original alignments: for each of the 20 sequences per alignment, the amino acid frequencies were computed and compared with the amino acid frequencies observed in the original alignments.

We also investigated the influence of the alignment size on the estimation of equilibrium frequencies (see Supplementary Fig. S2 that can be found in the Dryad data repository [doi:10.5061/dryad.7h66k]). We simulated 1000 alignments containing either 100 or 200 amino acids, with the same trees and sets of parameters as previously. This approach was motivated by two main reasons. First, it is not obvious whether NH–NS COaLA is able to generate accurate parameter estimates for short single-gene alignments. Second, in short alignments, some amino acids, especially rare amino acids such as tryptophan or cystein, may never be observed in any sequences. In such a case, the standard COA algorithm cannot be applied, since all elements of a column (here, the counts of a particular amino acid) are divided by its marginal sum. We devised a procedure to deal with such cases (see "Results" section and Supplementary Information). This procedure has proved to be efficient to avoid optimization problems.

To estimate the best model in terms of fitting data, either homogeneous or nonhomogeneous, BIC values (Schwarz 1978) were computed for each model (Felsenstein 2004; Ripplinger and Sullivan 2008) to penalize the number of parameters influencing the likelihood. A rooted tree is characterized by $2s-2$ internal branches, $s$ being the number of species. In the case of the NH–NS COaLA model, we count $k$ axis positions optimized per branch and at the root, and add the $\alpha$ parameter of the $\Gamma$ distribution, which results in the total number $K$ of parameters

$$K = k \times (2s-2) + k + 1.$$

The BIC value is computed as:

$$\mathrm{BIC} = -2 \times \ln L + K \times \ln(n),$$

where $\ln L$ is the optimal log-likelihood and $n$ is the alignment length. In this study, the LG (Le and Gascuel 2008) empirical exchangeability matrix does not add free parameters to the model. However, if a general time reversible (GTR) matrix is considered, 190 free exchangeabilities have to be taken into account in the total number of parameters. Moreover, it is worth noting that other statistical criteria for model selection may be employed. AIC (Akaike 1974) is one such criterion, which penalizes complex models less than does BIC ($\mathrm{AIC} = -2 \times \ln L + 2 \times K$). We chose to employ BIC because it was observed that AIC tends to favor models that are too parameterized with phylogenomic data sets (see "Results" section). We thus recommend the use of this criterion for model selection on large alignments. However, the situation is rather different on single-gene alignments, where BIC may penalize too strongly the more complex models in comparison with AIC (see "Results" section).

We note here that the NH–NS COaLA model used to estimate evolutionary parameters on simulated data sets is more parameter rich than the model used to simulate sequences, as in the latter several branches share the same substitution matrix (See "Materials and Methods" section). Although these simulation experiments therefore are a clear example of overparameterization, we believe they can provide valuable information regarding the accuracy of the COaLA model. One way to avoid overparameterization in this simulation setting would be to use the algorithms presented in Dutheil et al. (2012), which select the best branch-heterogeneous model on a fixed tree by finding the optimal partition of branches according to statistical criteria such as AIC or BIC. As the work of both Dutheil et al. (2012) and ours are based on the Bio++ libraries, the COaLA model can be easily incorporated to these programs to select the best configurations of axis position assignments over the tree.

## Biological Data Sets

### *Phylogenomic Alignments*

The COaLA model was tested on four previously published phylogenomic data sets (see below). For each data set, rate heterogeneity across sites was modeled with a discretized $\Gamma$ distribution with four categories (Yang 1994).

*Yeast data set*.—This data set is a concatenation of 106 genes belonging to eight yeast species (Rokas et al. 2003). This alignment contains 42 342 amino acids and the species tree presented in Figure 4 of the corresponding paper is used to estimate evolutionary parameters and compute the likelihood. The G+C content of third codon positions is heterogeneous among the eight species, ranging from 0.28 in *Candida albicans* to 0.45 in *Saccharomyces kluyveri*, possibly influencing the composition at the amino acid level. In line with this, 46%

of the pairwise Bowker tests performed on the protein concatenate are statistically significant (according to Holm correction for multiple comparisons).

*Archaea data set*.—These data are a concatenation of 72 protein-coding genes sampled in 35 archaeal species and 10 bacterial species (Groussin and Gouy 2011). We removed bacteria from the alignment, as well as the two uncultured thaumarchaeal species, for which only one protein sequence was present in the alignment. The final alignment of 9387 amino acids contains 33 archaeal species. We used the topology presented in figure 3 of Groussin and Gouy (2011) to determine the best evolutionary model with BppML. These sequences are compositionally highly heterogeneous since 86% (after correction for multiple tests) of the pairwise Bowker tests significantly rejected the stationarity or homogeneity hypotheses.

*Eocyte data set*.—Cox et al. (2008) used 45 genes to build a universal alignment of 5521 sites and 40 species. Using a nonhomogeneous model that allowed them to explore the space of tree topologies in the Bayesian framework, they obtained a topology called "eocyte" where Crenarchaea is the sister group of Eukaryotes. This topology was used in our analysis of their alignment. The compositional heterogeneity present in the data is strong, with 77% significant pairwise Bowker tests (after multiple tests correction).

*Three domains data set*.—Boussau et al. (2008) used 56 unicopy genes to build a universal alignment of 30 species. Because of a drastic selection of sites allowing only sites with less than 5% of gaps to remain in the final alignment, the total number of sites is rather small (3336 sites). We increased the size of the final alignment by using a less drastic site selection. Each individual gene alignment was realigned with Muscle v3.7 (Edgar 2004), internally used by Guidance v1.1 (Penn et al. 2010) with its default parameters. Guidance is a program allowing users to evaluate the reliability of alignments by taking into account the uncertainty of the guide tree used to align sequence positions with a bootstrap procedure. The resulting alignments were then treated by Gblocks (Castresana 2000) to eliminate ambiguous regions (default parameters with the authorization to conserve gap sites were used). The final gene alignments were concatenated and the sites with more than 50% of gaps were removed to eventually obtain an alignment of amino acids with 6269 sites.

### *Single Gene Alignments*

To evaluate both the ability to fit the data and the accuracy of ancestral sequence reconstruction on single-gene alignments with NH–NS COaLA in comparison with a homogeneous model, gene alignments were constructed from 24 methanogenic archaeal genomes (15 Methanococcales, 8 Methanobacteriales, and 1 Methanopyrale, see Supplementary Table S2). This data set presents two advantages: these species do not have extreme rates of evolution (Brochier-Armanet et al. 2011) and are adapted to different optimal growth temperatures (OGTs), leading to compositional variability (Groussin and Gouy 2011). All genome sequences were retrieved from GenBank. The software package SiLiX (Miele et al. 2011) was employed to cluster amino acid sequences into homologous gene families. Unicopy gene families containing at least 80% of the 24 species were conserved, leading to 535 gene families. Each family was further aligned with PRANK (Löytynoja and Goldman 2008) internally used by Guidance. The resulting alignments were then trimmed by Gblocks (Castresana 2000) with default parameters and the authorization to conserve gap sites. Phylogenetic trees were computed with PhyML (Guindon and Gascuel 2003) with a WAG+$\Gamma$(4) model (Yang 1994; Whelan and Goldman 2001). The trees were subsequently mid-point rooted and used with their corresponding alignments to run COaLA in a NH–NS fashion with a LG+COaLA[1]+$\Gamma$(4) model. From the ML estimates (model parameters and branch lengths), 535 alignments were simulated (one per set of ML estimates) with BppSeqGen (Dutheil and Boussau 2008). During simulations, ancestral sequences for each internal node were conserved and are henceforth referred to as "true" sequences. For each of the 535 simulated alignments, a model comparison was performed with the H–S LG+$F_{opt}$ and NH–NS LG+COaLA[1] models. With the ML estimates obtained with each model, ancestral sequences were computed with BppAncestor (Dutheil and Boussau 2008), with a marginal reconstruction (see Appendix). For each internal node, the ML pairwise distances between the homogeneously inferred sequence and the true sequence and between the nonhomogeneously inferred sequence and the true sequence were computed with the LG model (Le and Gascuel 2008).

## Results

### *Simulations*

*NH–COaLA accurately estimates ancestral amino acid frequencies*.—We verified that the compositional variance encountered in the simulated alignments was distributed as in biological data. For the first 1000 simulated alignments (out of 5000; see "Materials and Methods" section "Sequence simulations"), Supplementary Figure S1a shows that on average, the first three axes represent 53%, 23%, and 11% of the total variance, which is very similar to what can be observed in real sequences (Supplementary Fig. S1b–e and see below). This suggests that our simulated alignments have properties that are routinely encountered in biological data sets. When COaLA models were employed on these simulated data sets,

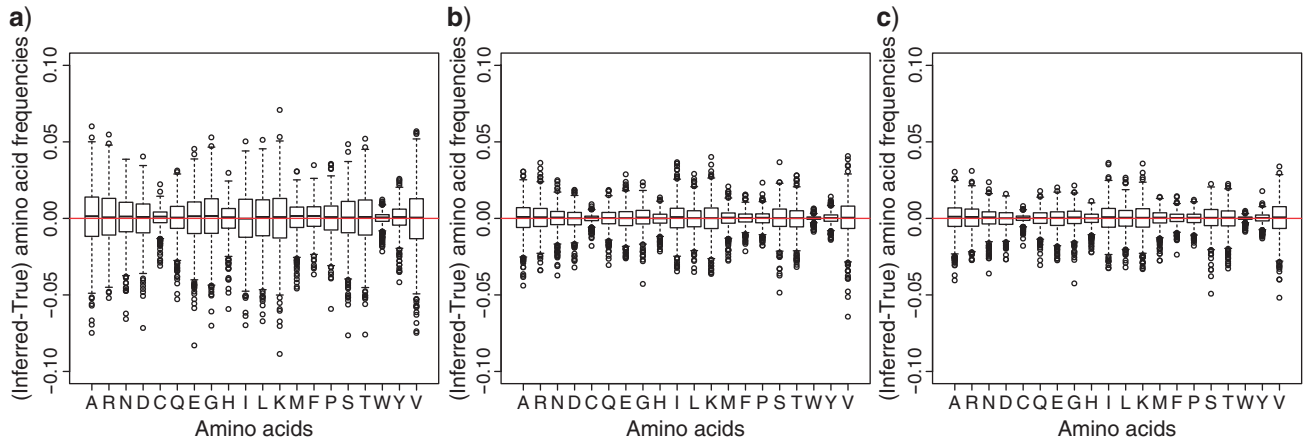



FIGURE 2. Accuracy of estimation of ancestral root amino acid frequencies. On the *y*-axes, the differences between inferred amino acids frequencies by ML and true amino acid frequencies used to simulate sequences are represented. a) Results obtained with the H–S LG+$F_{opt}$ model. b) Results obtained with the H–NS LG+COaLA[2] model. c) Results obtained with the NH–NS LG+COaLA[2] model.

two axis positions per branch were estimated, allowing to take into account, on average, about 75% of the variance (Supplementary Fig. S1a). Note that the NH–NS model with 19 free parameters per branch was not used in simulations as it generally takes too much time to converge. A comparison with the NH–NS COaLA model for calculation time and fit to data is provided with the analysis of real data (see below).

Figure 2 shows that for the first 1000 alignments, both the NH–NS and H–NS LG+COaLA[2] models outperform the H–S LG+F model when it comes to estimating ancestral root frequencies. The sums of the squared differences between true and inferred amino acid frequencies are equal to 4.38, 1.15, and 0.98 for the H–S, H–NS, and NH–NS models, respectively, with the NH–NS model exhibiting slightly better performances than the H–NS approach (Wilcoxon paired test, $P < 0.001$). Furthermore, we observed that for both rare (such as cysteine or tryptophan) or frequent amino acids (such as alanine), the NH–NS COaLA model remains the best ($P < 0.001$) at estimating ancestral frequencies at the root (the sums of the squared differences are, in the same order as before, 0.018, 0.0025, and 0.0022 for tryptophan and 0.398, 0.112, and 0.098 for alanine). This might be explained by the fact that COA is equally sensitive to deviations in rare amino acids as it is to deviations in frequent amino acids.

For the H–S, H–NS, and NH–NS approaches, we resimulated alignments from the parameters estimated by BppML to compare the ability of the different approaches to capture the evolutionary signal within the tree. We reasoned that if the model is able to correctly extract the signal from the data, sequences simulated from the ML parameter estimates should be close to the original sequences regarding their amino acid compositions. Thus, the amino acid frequencies of each simulated sequence were then computed and compared with the amino acid frequencies of the corresponding sequence from the original simulated alignment. Medians of squared differences of amino acid frequencies are presented in Figure 3. This figure highlights that the NH–NS approach better captures the evolution of compositional heterogeneities through time, as attested by the low-squared differences between simulated and expected amino acid frequencies.

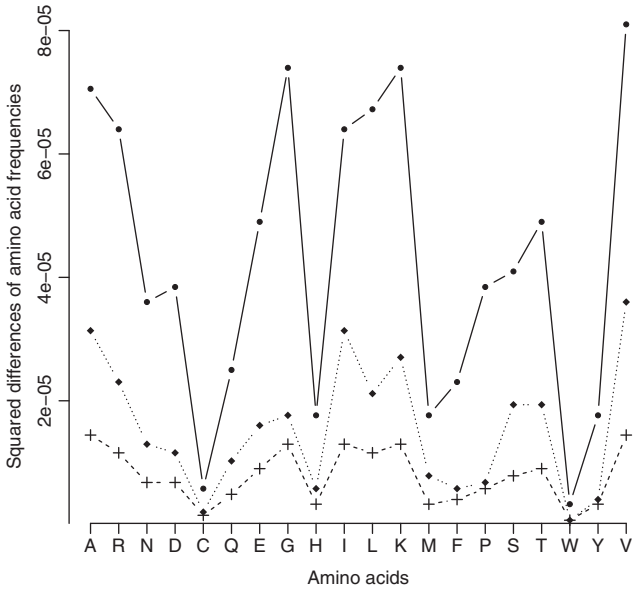



FIGURE 3. Accuracy of the phylogenetic signal capture. The H, H–NS, and NH–NS approaches are compared. For each of the original 1000 simulated alignments, parameters estimates were obtained with each one of the three approaches. From these estimates, new alignments were simulated with BppSeqGen. For each of the 20 sequences per alignment, the amino acid frequencies were computed and compared with the amino acid frequencies observed in the original alignments. The medians of squared differences for each amino acid frequency are represented. Solid line: H–S model. Dotted line: H–NS model. Dashed line: NH–NS model. The NH–NS approach is the best approach regarding the modeling of evolutionary processes and the capture of the phylogenetic signal present in the data.

*Influence of the size of the alignment on the estimation of ancestral amino acid frequencies.*—For short alignments, the standard NH–NS COaLA model might experience optimization problems for rare amino acids that may be totally absent in the alignment. We implemented a special procedure to deal with such cases (see "Materials and Methods" section and Supplementary Information for a full description) to avoid optimization issues.

The branch-wise NH–NS models can be expected to perform poorly with short alignments, because a large amount of data is needed to accurately optimize the equilibrium frequencies. Accordingly, Supplementary Figure S2 shows that the optimization of ancestral root frequencies for amino acid aligments with 100 sites is less accurate than what can be obtained with a H–S model: the sums of squared differences between the estimated frequencies and the true frequencies are equal to 7.64 and 6.72, respectively ($P < 0.001$). However, for amino acid aligments with 200 sites, NH–NS COaLA becomes better than a homogeneous model (4.55 and 5.27, respectively [$P < 0.001$], data not shown).

*NH–COaLA accurately estimates branch lengths.*—The ability of NH–NS COaLA to accurately estimate branch lengths was assessed (See Supplementary Information). Supplementary Figure S3 shows that NH–NS COaLA has similar performances to an H–S LG+$F_{obs}$ model without any bias.

*NH–COaLA efficiently fits data.*—The H–S, H–NS, and NH–NS sequence evolution models were compared using the BIC, which aims at identifying the best compromise between fit of the model to the data (likelihood) and small number of parameters. We used the 272 most heterogeneous alignments (out of 5000 simulations), whose compositional heterogeneity, measured by the fraction of statistically significant Bowker tests, is comparable with what can be observed in real data (See "Materials and Methods" section).

The NH–NS LG+COaLA model with one or two parameters per branch outperforms, according to the BIC, the H–S LG+$F_{opt}$ model in 53% and 70% of the 272 cases, respectively. Furthermore, the H–NS model is better than the H–S model only in 10% of the cases and is better than the NH–NS LG+COaLA model with one and two parameters per branch only in 11% and 5% of the cases, respectively. These results illustrate the excellent fit of nonhomogeneous evolutionary models to compositionally heterogeneous sequences.

*Model misspecifications.*—If one considers two Markovian transition probability matrices, $P_1 = e^{Q_1 l_1}$ and $P_2 = e^{Q_2 l_2}$, modeling the evolutionary process along two neighboring branches of length $l_1$ and $l_2$, the transition probability matrix $P'$ modeling evolution along the combined branch can be expressed as $P' = P_1 P_2$. In a recent article, Sumner et al. (2012b) demonstrated that the GTR model (Yang 2006), as well as several other substitution models in the context of DNA sequences, lacks closure under matrix multiplication. More precisely, if $P_1$ and $P_2$ are two GTR transition probability matrices with different exchangeabilities and/or equilibrium frequencies, their product $P'$ is not a GTR transition probability matrix, but belongs to a different model class. However, if $P_1$ and $P_2$ have identical exchangeabilities and equilibrium frequencies but differ by their branch lengths only, their product $P'$ is a GTR probability matrix.

These considerations have a direct bearing on our ability to infer evolutionary parameters. If one assumes that the data have been generated through a succession of GTR matrices that differ in their exchangeabilities and/or equilibrium frequencies along branches of the phylogeny, then a GTR-based model is bound to make some error, and a proper model to perform inference would be a model that has the closure property. In contrast, if one assumes that the data have been generated through a succession of GTR matrices that differ in their branch lengths only, then the closure property ensures that a H–S GTR-based model can correctly estimate the parameters of the model provided there is enough data.

It is of interest to determine whether the model considered here, a single empirical exchangeability matrix of the GTR-based LG model (Le and Gascuel 2008) with branch-wise equilibrium frequencies lacks closure under multiplication and is, as a result, affected by the type of misspecification studied in Sumner et al. (2012b).

To verify this point, and to quantify the amount of misspecification affecting our approach, we performed an experiment similar to Sumner et al. (2012a). These authors measured how much the nonclosure of the GTR model affects the estimation of transition probabilities for DNA sequences. In our case, two LG substitution matrices $P_1$ and $P_2$ were employed, with the equilibrium frequencies $\pi_1$ and $\pi_2$ of $Q_1$ and $Q_2$ drawn from the same Dirichlet distribution as presented above, modeling the succession of two independent substitution models along two successive branches. We then computed the product $P' = P_1 P_2$, with both $l_1$ and $l_2$ equal to 0.5 substitutions/site. Finally, we computed the equilibrium frequencies of another substitution matrix $\bar{P}$ with the same LG exchangeability matrix that minimized its distance to $P'$ using the Euclidean distance between matrices:

$$d(P', \bar{P}) = \sqrt{\sum_{i \neq j} (P'_{ij} - \bar{P}_{ij})^2}.$$

This distance measures the amount of misspecification caused by the nonclosure property of the model. If the minimization procedure finds equilibrium frequencies so that this distance is zero, the model has the desired closure property. If not, the model is nonclosed under multiplication and the distance reflects the amount of errors in the estimation of transition probabilities due to the nonclosure property. We ran 1000 simulations and, for each simulation, we measured both the average

percentage error and the average absolute difference between corresponding transition probabilities of $P'$ and $\bar{P}$. We observed that the mean distance $\hat{d}(P', \bar{P})$ is 0.02. Furthermore, the mean percentage error in transition probabilities is 5.0% and the mean absolute difference is $7 \times 10^{-4}$. These results show that, like the GTR model for nucleotide sequence evolution (Sumner et al. 2012a), our model of amino acid sequence evolution based on a fixed LG exchangeability matrix with optimized equilibrium frequencies lacks closure under multiplication. In both cases, it remains to be seen to what extent this creates a problem for evolutionary inference of parameters, phylogenetic trees, and ancestral sequences. It will be further interesting to study how H–S versus NH–NS models cope with such model misspecifications. Nonetheless, despite the nonclosure property of the model employed here, NH–NS COaLA brings strong benefit in terms of data fitting or inference of ancestral frequencies and sequences in comparison with the H–S model.

### *Tests on Phylogenomic Data Sets*

*COA of the observed frequencies.*—The concatenated alignments of yeast, archaea, and eocyte sequences are studied here (the fourth "Three domains" alignment is analyzed later). For each of these alignments, a matrix of observed amino acid frequencies was computed and used to compute a COA. For the yeast data set, the first and second axes account, respectively, for 63% and 32% of the total variance initially present in the data, meaning that the plane defined by the first two factors of the COA reflect 95% of the total compositional variance in the data (Supplementary Fig. S1b). In the eocyte data set, the first three axes account, respectively, for 46%, 24%, and 9% (Supplementary Fig. S1c), while their contribution is 43%, 28%, and 10% (Supplementary Fig. S1d), respectively, in the archaea data set. These variation axes are strongly linked to biological properties that influence the global amino acid composition of proteomes. We observed that the first axis of the COA highly correlates with the G+C content of third codon positions (GC3) of each yeast species ($r = -0.89$). In the eocyte data set, the first factor discriminates eukaryotic from archaeal/bacterial species. The second factor highly correlates with the genomic G+C content ($r = 0.9$) and the third factor is strongly linked to OGT ($r = 0.88$). Finally, in the archaeal data set, the first and second axes highly correlate with the genomic G+C content ($r = 0.74$) and the OGT ($r = 0.83$), as previously reported (Groussin and Gouy 2011).

*NH–COaLA fits the data better than homogeneous models.*—We applied the COaLA model to these biological data sets to estimate the ML values of branch lengths and evolutionary parameters. Table 1 summarizes the results. In all cases, according to the BIC, the NS COaLA model fits the sequence data better than the best homogeneous and stationary model ($LG+F_{opt}$). For the yeast data set, the H–NS model is the best model in terms of BIC values. It is interesting to note that the COaLA model, used in the homogeneous case with fewer parameters, provides a better fit than the classic $LG+F_{opt}$ model. Concerning archaea, the best model is the NH–NS LG+COaLA[3] model. However, only two axis positions per branch were necessary to best fit the eocyte data set. It is surprising to observe that in this case the $LG+F_{obs}$ model fits the data less well than the LG model, where the vector of equilibrium frequencies is the one empirically estimated by (Le and Gascuel 2008), on several biological data sets. The exact same final likelihood was also obtained using PhyML, which indicates that this unexpected result is not a problem specifically found by BppML. We hypothesize that this is because the observed frequencies are not ML estimates and potentially lead to worse likelihood scores. Finally, we found that using AIC instead of BIC for model selection (see "Materials and Methods" section) systematically leads to the choice of overparameterized models, illustrating the property of

TABLE 1. Assessing the fit to the data between several evolutionary models

| Data set | Process | Model | ln*L* | nbr Param | BIC |
|---|---|---|---|---|---|
| Yeast | H–S | LG | −299506.1 | 1 | 599022.9 |
| | | $LG+F_{obs}$ | −298702.5 | 1 | 597415.7 |
| | | $LG+F_{opt}$ | −298575.3 | 20 | 597363.7 |
| | | LG+COaLA[1] | −298667.9 | 2 | 597357.1 |
| | **H–NS** | **LG+COaLA[1]** | **−298 595.4** | **3** | **597 222.8** |
| | NH–NS | LG+F | −297621.7 | 286 | 598290.3 |
| | | LG+COaLA[1] | −298543.5 | 16 | 597257.5 |
| | | LG+COaLA[2] | −298505.3 | 31 | 597340.9 |
| | | LG+COaLA[3] | −298500.6 | 46 | 597491.3 |
| | | LG+COaLA[4] | −298491.7 | 61 | 597633.3 |
| | | LG+COaLA[5] | −298486.4 | 76 | 597782.5 |
| Eocyte | H–S | LG | −277967.3 | 1 | 555943.2 |
| | | $LG+F_{obs}$ | −278064.5 | 1 | 556137.6 |
| | | $LG+F_{opt}$ | −277444.0 | 20 | 555060.3 |
| | | LG+COaLA[1] | −277877.0 | 2 | 555771.2 |
| | H–NS | LG+COaLA[1] | −277695.3 | 3 | 555416.4 |
| | NH–NS | LG+F | −274279.4 | 1502 | 561501 |
| | | LG+COaLA[1] | −277263.8 | 80 | 555216.9 |
| | | **LG+COaLA[2]** | **−276 483.0** | **159** | **554 336** |
| | | LG+COaLA[3] | −276253.3 | 238 | 554557.3 |
| | | LG+COaLA[4] | −276090.3 | 317 | 554912 |
| | | LG+COaLA[5] | −275946.5 | 396 | 555305.1 |
| Archaea | H–S | LG | −340369.1 | 1 | 680747.3 |
| | | $LG+F_{obs}$ | −340047.3 | 1 | 680103.7 |
| | | $LG+F_{opt}$ | −339217.9 | 20 | 678618.7 |
| | | LG+COaLA[1] | −339887.8 | 2 | 679793.9 |
| | H–NS | LG+COaLA[1] | −339865.7 | 3 | 679758.8 |
| | NH–NS | LG+F | — | 1236 | — |
| | | LG+COaLA[1] | −338985.4 | 66 | 678574.5 |
| | | LG+COaLA[2] | −338237.7 | 131 | 677673.7 |
| | | **LG+COaLA[3]** | **−337 932.3** | **196** | **677 657.4** |
| | | LG+COaLA[4] | −337721.0 | 261 | 677829.4 |
| | | LG+COaLA[5] | −337541.1 | 326 | 678064.1 |

Bold lines highlight the best model according to the BIC.

BIC to more heavily penalize parameter-rich models. For instance, with the archaea data set, AIC selects the NH–NS LG+COaLA[7] model, where the seventh axis of the COA only represents 1.4% of the total compositional variance of the data.

With respect to the number of parameters involved, the COaLA model strongly reduces the dimension of the evolutionary model. Consequently, COaLA is fast and saves a large amount of computing time: with the yeast data set containing eight species, 5 h 32 min were necessary to compute the likelihood with 19 equilibrium frequencies per branch in comparison with 2 h 38 m for the NH–NS COaLA[1] model and with 16 min 14 s for the H–NS COaLA[1] model. Concerning the eocyte data set, the model with 19 equilibrium frequencies per branch required about 522 h of calculation to converge to the ML optimum. Comparatively, the best COaLA model only required about 40 h of calculation. For the two other data sets (archaea and three domains), we cannot provide a precise comparison as the 19 equilibrium frequencies per branch model was stopped after 1 month of calculation before reaching the ML optimum. The best COaLA models used about 26 and 18 h of calculation, respectively, with a very stringent threshold of $10^{-6}$ below which convergence is accepted.

### *Tests on Single Gene Data Sets*

*NH–COaLA is overparameterized for single-gene alignments.*—From the 24 methanogenic archaeal genomes, we built all homologous gene families (see "Materials and Methods" section) and conserved the unicopy and nearly universal families, leading to 535 genes. For each of these gene families and their corresponding ML phylogenetic trees (see "Materials and Methods" section), we compared the performance of the NH–NS LG+COaLA model with the best H–S model (LG+$F_{opt}$) regarding the fit to the data. Only in 19 cases did the NH–NS LG+COaLA[1] model with the optimization of one axis position per branch outperform the homogeneous model, according to the BIC criterion. However, the NH–NS LG+COaLA[1] model outperformed the homogeneous model in 172 cases according to AIC. Overall, these results indicate that with small single-gene alignments, COaLA may model the evolutionary process more accurately than homogeneous models but is generally overparameterized, calling for future improvements (see "Discussion" section). However, in all estimations, we did not observe unconventional frequencies for rare amino acids, showing that the way COaLA copes with the problem of completely absent amino acids (see "Materials and Methods" section and Supplementary Information) is robust.

*NH–COaLA reconstructs ancestral sequences more accurately.*—In studies using ancestral sequence reconstruction and resurrection, major biological conclusions can sometimes rely on one or few amino acid differences between ancient or between extant and ancient proteins (Finnigan et al. 2012; Huang et al. 2012). However, these substitutions may differ depending on the model employed. Here, we attempt to test whether the NH–NS COaLA model can lead to better ancestral sequence reconstruction, at the single gene level, in comparison with the H–S LG+$F_{opt}$ model.

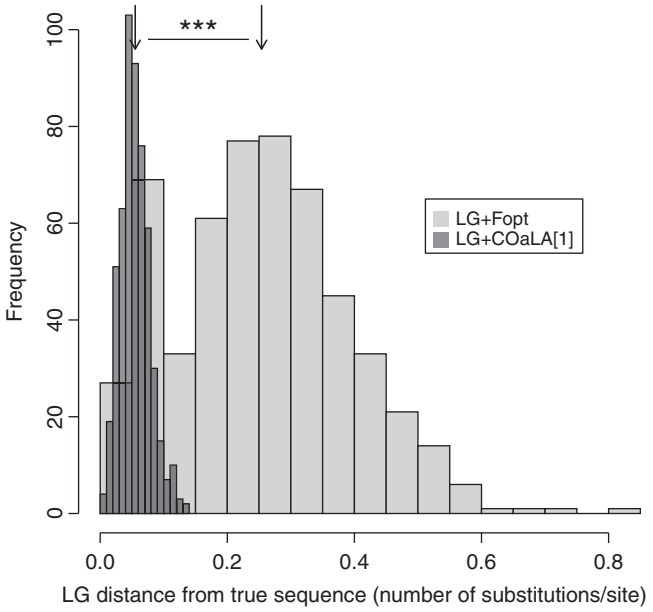



FIGURE 4. Accuracy of the ancestral sequence reconstruction. With the 535 simulations of single-gene alignments (see "Materials and Methods" section), ancestral sequence reconstruction was performed with a H–S model (LG+$F_{opt}$) and with a NH–NS model (LG+COaLA[1]). For all ancestral sequences, a LG distance was computed between the inferred and the true sequences recorded during the simulation procedure. For each of the 535 cases, the mean LG distance was calculated and the distribution of means is represented in light and dark gray for the LG+$F_{obs}$ and LG+COaLA[1] models, respectively. The mean of the distributions (black arrows) are 0.25 and 0.06, respectively ($P < 0.001$).

We simulated the evolution of 535 gene families using the parameter values (sequence length, tree shape, branch lengths, and amino acid equilibrium frequencies) given by the 535 alignments of methanogenic archaea described above. We ran NH–NS COaLA[1] and H–S LG+$F_{opt}$ on these 535 alignments simulated in a nonhomogeneous fashion and then reconstructed the ancestral sequences with BppAncestor for all internal nodes (see Appendix). These inferred sequences were then compared with a LG distance (computed by ML) to their true corresponding sequences recorded during the simulation procedure. For each of the 535 simulations, we computed the average distance for all nodes. Figure 4 shows the distribution of the 535 mean LG distances for the two models. First, the NH–NS COaLA model outperforms the best H–S model (LG+$F_{opt}$) regarding the accuracy of ancestral sequence reconstruction. Second, the mean of the distribution of the LG+$F_{opt}$ model is 0.25 substitution/site, meaning that every four sites on average, an amino acid difference

exists between the inferred and the true sequence with the H–S approach. In contrast, the mean distance is reduced to 0.06 with the NH–NS approach of ancestral sequence reconstruction.

### *NH–COaLA Confirms the Mesophilic State of the Last Universal Common Ancestor*

This section is focused on the three-domains data set used by Boussau et al. (2008) to study the early pattern of adaptation to temperature. Given the results presented above concerning the performances of COaLA, we used the NH–NS approach to infer the ancestral environmental temperatures over the universal Tree of Life. We first demonstrate that COaLA accurately fits the data with the concatenate alignment. Finally, we confirm the results regarding the early adaptation to environmental temperature obtained by Boussau et al. (2008) with a different NH–NS model.

*Capturing the nonhomogeneity of the data.*—We first determined the best model. Supplementary Table S1 shows that the NH–NS LG+COaLA[2] model better fits the data than the other models according to BIC. Given the ML estimates of the evolutionary parameters obtained with this model, 200 simulated alignments of similar size as the original alignment were produced to check the capability of COaLA to capture the heterogeneity present in the data. On average, 35% of the Bowker pairwise tests were significant after the Holm–Bonferroni correction, in comparison with 38% significant tests observed on the original alignment. Consequently, according to this measure, 92% of the original compositional heterogeneity is captured by the model, even though only two parameters per branch are used. When the NH–NS LG+COaLA[3] model is used, thereby optimizing three axis positions per branch instead of two, simulated alignments have on average a higher level of heterogeneity than the original data (41% vs. 38%, respectively). The BIC criterion is therefore conservative and favors a model with fewer parameters, even if it does not capture all the heterogeneity in the data.

*The COaLA model confirms the early pattern of adaptation to temperature.*—Boussau et al. (2008) proposed that the Last Universal Common Ancestor (LUCA) lived in a mesophilic environment in opposition to its two descendants, inferred as being thermophilic organisms. They used the strong relationship that exists between either the G+C content in rRNAs or the amino acid contents in proteins and the OGT of bacteria and archaea. This relation allows constructing molecular thermometers (Galtier and Lobry 1997; Boussau et al. 2008; Groussin and Gouy 2011) that give estimates of environmental temperatures from ancestral amino acid or nucleotide compositions. With nonhomogeneous models of evolution, Boussau et al. (2008) inferred the ancestral compositions for all nodes of a universal tree and estimated the corresponding OGTs with the molecular thermometers. For proteins, these inferences were realized with the NH–NS CAT-BP model (Blanquart and Lartillot 2008) in the Bayesian framework. Since COaLA and CAT-BP are implemented in different frameworks and model differently the nonhomogeneity of the evolutionary process, it is interesting to determine whether they give similar estimations of ancestral equilibrium frequencies and OGTs. With CAT-BP, Boussau et al. (2008) inferred that LUCA lived at 20°C [0–37°C] and the ancestors of bacteria and archaea+eukarya at 69°C [64–75°C], and 55°C [45–65°C], respectively. NH–NS COaLA also recovered a signal for a parallel adaptation to high temperatures from LUCA to its two descendants (*Wilcoxon test*, $P < 0.001$), with estimates that are very close to the ones obtained with CAT-BP. Thus, the ancestral OGTs are 34°C [24–44°C], 69°C [64–76°C], and 57°C [46–70°C] for LUCA, the ancestor of bacteria and the ancestor of archaea+eukarya, respectively. The 95% confidence intervals were computed with a nonparameteric bootstrap procedure. It is interesting to observe that with two different approaches, the COaLA and CAT-BP models converge toward a similar phylogenetic signal for the evolution of amino acid frequencies during early life and quantitatively similar estimates of ancestral compositions and temperatures.

## Discussion

When phylogenetic data are consistent with the assumption of compositional homogeneity, homogeneous models are often more suited for model-based phylogenetic analyses than nonhomogeneous models. In these cases, it is advisable to use a H–S model where the 20 equilibrium frequencies are fitted to the data by likelihood optimization (i.e., use the "$+F_{opt}$" model). Indeed, for all biological data sets investigated here, the gains of likelihood were significant when the 19 free equilibrium frequencies were estimated by ML. To our knowledge, BppML is the only phylogenetic program capable of generating ML estimates of the equilibrium amino acid frequencies (most other phylogenetic programs that we have checked appear to assume that the equilibrium amino acid frequencies are either equal to the equilibrium frequencies of the empirical model or to the observed amino acid frequencies).

Following Galtier and Gouy (1998), Galtier et al. (1999), Foster (2004), Jermiin et al. (2004), Gowri-Shankar and Rattray (2007), Blanquart and Lartillot (2008), and Boussau et al. (2008), we confirm the importance of using a nonhomogeneous and nonstationary model to estimate evolutionary parameters when compositional heterogeneity is present in the data. The COaLA model appears to be very efficient for the estimation of ancestral frequencies and to better fit heterogeneous data than classic NH or H models.

COaLA is flexible in the sense that it may be employed either as an H–S, H–NS, or NH–NS model. In the NH–NS approach, COaLA is a branch-wise heterogeneous model that assumes that (i) each branch is characterized by its own set of equilibrium frequencies and (ii) all branches share a common exchangeability matrix. Contrarily to Galtier and Gouy (1998) who used G+C equilibrium content as branch-wise variable irrespective of the nucleotide sequence data set under study, for each protein data set, the COaLA model constructs the branch-wise variables that summarize most of the variance in the data set under study. Therefore, the nature of the branch-wise variables differs among data sets. Previous authors mentioned the possibility that such branch-wise models may be overparameterized (Foster 2004; Blanquart and Lartillot 2006), as they assume that, at each speciation node, equilibrium frequencies evolve toward different positions in the space of frequencies. COaLA performs an efficient reduction of the parameter space used to optimize branch stationary frequencies. In all phylogenomic experiments, we showed that the model is very efficient at estimating evolutionary parameters such as ancestral frequencies or branch lengths. Even with rather small (5000 sites) phylogenomic data sets in the simulation experiments, and when the heterogeneity is similar to what one can observe with real data, the model is on average better than a homogeneous model. Overparameterization by the branch-wise approach in comparison with a homogeneous approach was detected in only 30% of the cases according to BIC with simulation experiments of sequence alignments having levels of compositional heterogeneity comparable with empirical data. With real data, three out of the four phylogenomic data sets were more efficiently fitted by the NH–NS branch-wise model than by other models. With more and more biological data coming from many and diverse sequencing projects, the data set sizes should increase as well. We observed that large, concatenated data sets are less frequently overparameterized by NH–NS models than single-gene data sets. This suggests that overparameterization may become less of an issue for data sets of increasing size.

Besides, we also demonstrated that the use of branch-heterogeneous models is crucial to infer accurate ancestral sequences. This result may be especially relevant for protein resurrection experiments where the accuracy of ancestral sequence reconstruction is crucial. Consequently, we strongly recommend the use of nonhomogeneous models for such studies when homologous sequences are observed to be compositionally different.

In many studies, NH–NS models were proved to better capture the evolutionary signal and to improve our knowledge concerning various biological questions (Herbeck et al. 2005; Nabholz et al. 2011; Boussau and Gouy 2012). Using NH–NS protein models in the Bayesian framework, Boussau et al. (2008) proposed that LUCA was a mesophilic organism and that its two descendants independently adapted to higher temperatures. This nonparsimonious scenario raised questions about possible biases in the models used to infer ancestral compositions. In their study, Boussau et al. (2008) extensively tested that their prediction was not the result of a bias in the model employed. They showed that this parallel adaptation to high temperatures was also recovered with different universal topologies and in the presence or absence of Eukaryotes. In this study, we confirmed this evolutionary pattern of adaptation to OGT with NH–NS COaLA using a ML rather than a Bayesian approach.

The COaLA model presented here is implemented in the ML framework but could be easily defined in a Bayesian context. Further theoretical work might improve the fit of the COaLA model to protein sequences. First, to further reduce the number of free parameters, a discretized version of the model could be developed. As already shown in Boussau and Gouy (2006) for nucleotide sequences, the model could propose a subset of fixed or optimized axis positions per branch, making it less flexible. For each branch, the best of the possible axis positions would be retained and could be used to compute the likelihood. This procedure could be especially relevant for single-gene alignments, where overparameterization was detected in this study. Second, the time-wise nonhomogeneity of the model could be extended with site-wise nonhomogeneity. Currently, the CAT-BP model (Blanquart and Lartillot 2008), in the Bayesian context, is able to combine the modeling of compositional variations both over time and over sites. However, the major drawback of this model is its huge computational cost, underlining the need for a more efficient model. To model the variation of evolutionary processes among sites, several approaches are already available, such as the mixture models implemented by Le et al. (2008b), or the empirical profile mixture models developed by Le et al. (2008a) (analogous to the CAT model [Lartillot and Philippe 2004] available in the Bayesian framework). Therefore, COaLA could be extended to the use of mixture models for which the equilibrium frequencies of each category would be modulated by the equilibrium frequencies of the branch under consideration. With such site and branch heterogeneity, COaLA would better take into account the variation of substitution processes depending on the localization of the residue in the protein 3D structure or depending on amino acid biochemical properties.

## Supplementary Material

Supplementary information can be found in the Dryad data repository at http://datadryad.org, doi:10.5061/dryad.7h66k.

## Funding

This work is a contribution to the project Ancestrome supported by the Agence Nationale de la Recherche (ANR-10-BINF-01-01).

## Acknowledgments

Sincere thanks to Lars Jermiin, Greg Fournier, and two anonymous reviewers for their suggestions and comments which greatly improved this article. The authors are also particularly thankful to Julien Dutheil, Anne-Béatrice Dufour, Jean Thioulouse, and all other members of the Bioinformatics and Evolutionary Genomics team for suggestions and fruitful discussions.

## Appendix

### *Correspondence Analysis*

We summarize here the principles used to compute a COA, which is necessary in order to understand the COaLA model. For more details about the specific properties of a COA, see (Greenacre 1984).

Let $I$ and $J$ be the number of rows and columns, respectively, of the matrix $\mathbf{N}_{I\times J}$ with elements $n_{ij}$, where $n_{ij}$ corresponds to the observed frequency of amino acid $j$ in sequence $i$, $I$ corresponds to the number of sequences in the alignment ($i=1,\ldots,I$), and $J$ corresponds to the number of different amino acids in the alignment ($j=1,\ldots,J$). Let $n_{i\bullet}$ and $n_{\bullet j}$ be the sum of the $i$th row and $j$th column, respectively, and $n$ denotes the total sum of $\mathbf{N}_{I\times J}$:

$$n_{i\bullet}=\sum_{j=1}^{J}n_{ij};\quad n_{\bullet j}=\sum_{i=1}^{I}n_{ij};\quad n=\sum_{i=1}^{I}\sum_{j=1}^{J}n_{ij}.$$

The matrix $\mathbf{P}_{I\times J}$ of relative frequencies $p_{ij}$ is then derived, so that:

$$p_{ij}=\frac{n_{ij}}{n};\quad p_{i\bullet}=\frac{n_{i\bullet}}{n};\quad p_{\bullet j}=\frac{n_{\bullet j}}{n},$$

where $p_{i\bullet}$ and $p_{\bullet j}$ represent the row and column weights, respectively.

Let $\mathbf{D}_{I\times I}$ and $\mathbf{D}_{J\times J}$ be the following diagonal matrices:

$$\mathbf{D}_{I\times I}=\mathrm{diag}(p_{1\bullet},\ldots,p_{I\bullet});\quad \mathbf{D}_{J\times J}=\mathrm{diag}(p_{\bullet 1},\ldots,p_{\bullet J}).$$

The matrix $\mathbf{Z}_{I\times J}$ is then computed:

$$\mathbf{Z}_{I\times J}=\mathbf{D}_{I\times I}^{-1}\mathbf{P}_{I\times J}\mathbf{D}_{J\times J}^{-1}-\mathbf{1}_{I\times J}$$

with:

$$\mathbf{D}_{I\times I}^{-1}=\mathrm{diag}\left(\frac{1}{p_{1\bullet}},\ldots,\frac{1}{p_{I\bullet}}\right);\quad \mathbf{D}_{J\times J}^{-1}=\mathrm{diag}\left(\frac{1}{p_{\bullet 1}},\ldots,\frac{1}{p_{\bullet J}}\right).$$

The general term of $\mathbf{Z}_{I\times J}$ is

$$z_{ij}=\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}-1=\frac{p_{ij}-p_{i\bullet}p_{\bullet j}}{p_{i\bullet}p_{\bullet j}}.$$

$\mathbf{Z}_{I\times J}$ is the table analyzed by the COA and represents the distance between expected under independence and observed frequencies.

To obtain the eigen elements of the COA, the matrix $\mathbf{H}$ containing the $\chi^2$ distances is computed:

$$\mathbf{H}_{J\times J}=\mathbf{D}_{J\times J}^{-1/2}\mathbf{Z}_{J\times I}^{T}\mathbf{D}_{I\times I}\mathbf{Z}_{I\times J}\mathbf{D}_{J\times J}^{-1/2},$$

with

$$\mathbf{D}_{J\times J}^{-1/2}=\mathrm{diag}\left(\frac{1}{\sqrt{p_{\bullet 1}}},\ldots,\frac{1}{\sqrt{p_{\bullet J}}}\right).$$

Next, $\mathbf{H}_{J\times J}$ is diagonalized to determine its eigenvalues and eigenvectors. The $k$ first eigenvalues in decreasing order are conserved and stored in $\Lambda_k$. The $k$ first associated eigenvectors, which are orthonormal, are stored as columns in $\mathbf{U}_{J\times k}$. $\mathbf{U}_{J\times k}$ possesses $J$ rows, $k$ columns, and verifies $\mathbf{U}_{k\times J}^{T}\mathbf{U}_{J\times k}=\mathbf{I}_{k\times k}$.

The row coordinates are computed with:

$$\mathbf{R}_{I\times k}=\mathbf{Z}_{I\times J}\mathbf{D}_{J\times J}^{1/2}\mathbf{U}_{J\times k}.$$

The columns of $\mathbf{R}_{I\times k}$ are the row coordinates. The columns' coordinates may also be computed:

$$\mathbf{C}_{J\times k}=\mathbf{D}_{J\times J}^{-1/2}\mathbf{U}_{J\times k}\Lambda_{k\times k}^{1/2}.$$

The columns of $\mathbf{C}_{J\times k}$ represent the column coordinates.

Once the COA is computed from a particular set of species, it may be useful to add a new row containing a set of values, where observed amino acid frequencies coming from another species. Thus, this vector of frequencies ($\mathbf{F}_{1\times J}$) defines a point in the space of the row profiles and it is possible to represent that point in the new space by projecting the point onto the space. To do so, the coordinates of the new vector in the new space can be calculated:

$$\mathbf{L}_{F}=\mathbf{F}_{1\times J}\mathbf{D}_{J\times J}^{1/2}\mathbf{U}_{J\times k}.$$

Conversely, from a set of row coordinates $\mathbf{L}_{F'}^{'}$, one can calculate a corresponding set of absolute frequencies $\mathbf{F}_{1\times J}^{'}$ in the original space using the matrix of column coordinates and accounting for the column weights (row weights are always equal to 1):

$$\mathbf{F}_{1\times J}^{'}=(\mathbf{L}_{F'}^{'}\mathbf{C}_{k\times J}^{T}+1)\mathbf{D}_{J\times J}.$$

Using this relation, from any set of coordinates in the new space, one can generate its corresponding set of frequencies in the original space of species profiles. It is worthwhile to note that one coordinate along the first axis of most variance is enough to propose a set of corresponding frequencies.

### *Ancestral Sequence Reconstruction*

We describe here how ancestral sequences are computed with a marginal reconstruction (Yang et al. 1995), either with a homogeneous or branch-heterogeneous model. In the following, we refer to the notations of figure 1 of Boussau and Gouy (2006). The BppAncestor program was used to compute for each site

and each inner node the posterior probabilities of each amino acid. The amino acid having the highest posterior probability is then retained in the ancestral sequence.

Consider the inner node $C$ in figure 1 of (Boussau and Gouy 2006). The marginal posterior probability of the state $\mathbf{v}$ is

$$P(C=\mathbf{v})=\frac{P(\mathbf{Data},C=\mathbf{v})}{P(\mathbf{Data})}.$$

where $P(\mathbf{Data})$ is the total likelihood $\mathbf{L}$ of the site. Using the upper conditional likelihoods introduced by Boussau and Gouy (2006), the joint probability of the data and having the state $\mathbf{v}$ at node $C$ is

$$P(\mathbf{Data},C=\mathbf{v})=\sum_y L_{s,\mathrm{Upp(UC)}}(U=y) \times P_{y\mathbf{v}}(l_C)\times L_{s,\mathrm{Low(UC)}}(C=\mathbf{v}),$$

where

- $L_{s,\mathrm{Low(UC)}}(C=\mathbf{v})$ is the lower conditional probability of having $\mathbf{v}$ at node $C$.
- $P_{y\mathbf{v}}(l_C)$ is the transition probability for a state $y$ to be substituted to $\mathbf{v}$ along a branch of length $l_C$
- $L_{s,\mathrm{Upp(UC)}}(U=y)$ is the upper conditional likelihood of having the state $y$ at the parent node $U$.

$L_{s,\mathrm{Upp(UC)}}(U=y)$ can be seen as the joint probability of the data excluding the part under node $C$ and having state $y$ at node $U$. It is recursively defined (Boussau and Gouy 2006) by

$$L_{s,\mathrm{Upp(UC)}}(U=y)=\left[\sum_x P_{xy}\times L_{s,\mathrm{Upp(RU)}}(R=x)\right]\left[\sum_q P_{yq}\times L_{s,\mathrm{Low(UB)}}(B=q)\right].$$

Thus, as mentionned in the "Materials and Methods" section of Boussau et al. (2008):

$$P(C=\mathbf{v})=\frac{P(\mathbf{Data},C=\mathbf{v})}{\mathbf{L}}=\frac{\sum_y L_{s,\mathrm{Upp(UC)}}(U=y)\times P_{y\mathbf{v}}(l_C)\times L_{s,\mathrm{Low(UC)}}(C=\mathbf{v})}{\mathbf{L}}$$

## References


Ababneh F., Jermiin L.S., Ma C., Robinson J. 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics 22:1225–1231.

Adachi J., Hasegawa M. 1996. MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. Comput. Sci. Monogr. 28:1–150.

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans. Autom. Contr. ACM 19:716–723.

Blanquart S., Lartillot N. 2006. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. Mol. Biol. Evol. 23:2058–2071.

Blanquart S., Lartillot N. 2008. A site- and time-heterogeneous model of amino acid replacement. Mol. Biol. Evol. 25:842–858.

Bollback J.P. 2002. Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. 19:1171–1180.

Boussau B., Blanquart S., Necsulea A., Lartillot N., Gouy M. 2008. Parallel adaptation to high temperature in the archaean eon. Nature 456:942–945.

Boussau B., Gouy M. 2006. Efficient likelihood computations with nonreversible models of evolution. Syst. Biol. 55:756–768.

Boussau B., Gouy M. 2012. What genomes have to say about the evolution of the Earth. Gondwana Res. 21:483–494.

Bowker A.H. 1948. A test for symmetry in contingency tables. J. Am. Stat. Assoc. 43:572–574.

Brochier-Armanet C., Forterre P., Gribaldo S. 2011. Phylogeny and evolution of the Archaea: one hundred genomes later. Curr. Opion. Microbiol. 14:274–281.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17:540–552.

Cox C.J., Foster P.G., Hirt R.P., Harris S.R., Embley T.M. 2008. The archaebacterial origin of eukaryotes. Proc. Natl Acad. Sci. U. S. A. 105:20356–20361.

Delsuc F., Brinkmann H., Chourrout D., Philippe H. 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439:965–968.

Douzery E.J.P., Snell E.A., Bapteste E., Delsuc F., Philippe H. 2004. The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? Proc. Natl Acad. Sci. U. S. A. 101:15386–15391.

Dutheil J., Boussau B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. BMC Evol. Biol. 8:255.

Dutheil J., Gaillard S., Bazin E., Glémin S., Ranwez V., Galtier N., Belkhir K. 2006. Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. BMC Bioinform. 7:188.

Dutheil J.Y., Galtier N., Romiguier J., Douzery E.J.P., Ranwez V., Boussau B. 2012. Efficient selection of branch-specific models of sequence evolution. Mol. Biol. Evol. 29:1861–1874.

Edgar R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.

Felsenstein J., editor. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Finnigan G.C., Hanson-Smith V., Stevens T.H., Thornton J.W. 2012. Evolution of increased complexity in a molecular machine. Nature 481:360–364.

Foster P.G. 2004. Modeling compositional heterogeneity. Syst. Biol. 53:485–495.

Galtier N., Gouy M. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proc. Natl Acad. Sci. U. S. A. 92: 11317–11321.

Galtier N., Gouy M. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol. 15:871–879.

Galtier N., Lobry J.R. 1997. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J. Mol. Evol. 44:632–636.

Galtier N., Tourasse N., Gouy M. 1999. A nonhyperthermophilic common ancestor to extant life forms. Science 283:220–221.

Gaucher E.A., Govindarajan S., Ganesh O.K. 2008. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. Nature 451:704–708.

Gowri-Shankar V., Rattray M. 2007. A reversible jump method for Bayesian phylogenetic inference with a nonhomogeneous substitution model. Mol. Biol. Evol. 24:1286–1299.

Greenacre M. 1984. Theory and applications of correspondence analysis. London: Academic Press.

Groussin M., Gouy M. 2011. Adaptation to environmental temperature is a major determinant of molecular evolutionary rates in Archaea. Mol. Biol. Evol. 28:2661–2674.
Guindon S., Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704.
Harms M.J., Thornton J.W. 2010. Analyzing protein structure and function using ancestral gene reconstruction. Curr. Opin. Struct. Biol. 20:360–366.
Hasegawa M., Hashimoto T. 1993. Ribosomal RNA trees misleading? Nature 361:23.
Herbeck J.T., Degnan P.H., Wernegreen J.J. 2005. Nonhomogeneous model of sequence evolution indicates independent origins of primary endosymbionts within the Enterobacteriales (Gamma-Proteobacteria). Mol. Biol. Evol. 22:520–532.
Ho S.Y.W., Jermiin L.S. 2004. Tracing the decay of the historical signal in biological sequence data. Syst. Biol. 53:623–637.
Hobbs J.K., Shepherd C., Saul D.J., Demetras N.J., Haaning S., Monk C.R., Daniel R.M., Arcus V.L. 2011. On the origin and evolution of thermophily: reconstruction of functional precambrian enzymes from ancestors of Bacillus. Mol. Biol. Evol. 29:825–835.
Holm S. 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6:65–70.
Huang R., Hippauf F., Rohrbeck D., Haustein M., Wenke K., Feike J., Sorrelle N., Piechulla B., Barkman T.J. 2012. Enzyme functional evolution through improved catalysis of ancestrally nonpreferred substrates. Proc. Natl Acad. Sci. U. S. A. 109:2966–2971.
Huelsenbeck J.P., Ronquist F., Nielsen R., Bollback J.P. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314.
Jayaswal V., Ababneh F., Jermiin L.S., Robinson J. 2011a. Reducing model complexity of the general Markov model of evolution. Mol. Biol. Evol. 28:3045–3059.
Jayaswal V., Jermiin L.S., Poladian L., Robinson J. 2011b. Two stationary nonhomogeneous Markov models of nucleotide sequence evolution. Syst. Biol. 60:74–86.
Jayaswal V., Jermiin L.S., Robinson J. 2005. Estimation of phylogeny using a general Markov model. Evol. Bioinform. Online 1:62–80.
Jayaswal V., Robinson J., Jermiin L. 2007. Estimation of phylogeny and invariant sites under the general Markov model of nucleotide sequence evolution. Syst. Biol. 56:155–162.
Jermiin L.S., Ho S.Y.W., Ababneh F., Robinson J., Larkum A.W.D. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst. Biol. 53:638–643.
Jermiin L.S., Jayaswal V., Ababneh F., Robinson J. 2008. Phylogenetic model evaluation. In: Keith J., editor. Bioinformatics—Volume I: data, sequences analysis and evolution. Totowa (NJ): Humana Press. p. 331–363.
Jones D.T., Taylor W.R., Thornton J.M. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8:275–282.
Lake J.A. 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc. Natl Acad. Sci. U. S. A. 91:1455–1459.
Lartillot N., Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino acid replacement process. Mol. Biol. Evol. 21:1095–2004.
Le S.Q., Gascuel O. 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25:1307–1320.
Le S.Q., Gascuel O., Lartillot N. 2008a. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics 24: 2317–2323.
Le S.Q., Lartillot N., Gascuel O. 2008b. Phylogenetic mixture models for proteins. Phil. Trans. R. Soc. Lond. B 363:3965–3976.
Lockhart P.J., Steel M.A., Hendy M.D., Penny D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11:605–612.
Löytynoja A., Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science 320:1632–1635.
Miele V., Penel S., Duret L. 2011. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinform. 12:116.
Nabholz B., Künstner A., Wang R., Jarvis E.D., Ellegren H. 2011. Dynamic evolution of base composition: causes and consequences in avian phylogenomics. Mol. Biol. Evol. 28:2197–2210.
Penn O., Privman E., Landan G., Graur D., Pupko T. 2010. An alignment confidence score capturing robustness to guide-tree uncertainty. Mol. Biol. Evol. 27:1759–1767.
Philippe H., Brinkmann H., Lavrov D.V., Littlewood D.T.J., Manuel M., Wörheide G., Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602.
Posada D. 2008. jModelTest: phylogenetic model averaging. Mol. Biol. Evol. 25:1253–1256.
Posada D., Crandall K.A. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817–818.
Ripplinger J., Sullivan J. 2008. Does choice in model selection affect maximum likelihood analysis? Syst. Biol. 57:76–85.
Rokas A., Williams B.L., King N., Carroll S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.
Schwarz G. 1978. Estimating the dimension of a model. Ann. Statist. 6:461–464.
Steel M. 2005. Should phylogenetic models be trying to “fit an elephant”? Trends Genet. 21:307–309.
Sumner J., Jarvis P., Fernandez-Sanchez J., Kaine B., Woodhams M., Holland B. 2012a. Is the general time-reversible model bad for molecular phylogenetics? Syst. Biol. 61:1069–1074.
Sumner J.G., Fernández-Sánchez J., Jarvis P. 2012b. Lie Markov models. J. Theor. Biol. 298:16–31.
Tamura K., Kumar S. 2002. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. Mol. Biol. Evol. 19:1727–1736.
Thioulouse J., Chessel D., Dolédec S., Olivier J.-M.. 1997. ADE-4: a multivariate analysis and graphical display software. Statist. Comput. 7:75–83.
Wertheim J.O., Sanderson M.J., Worobey M., Bjork A. 2010. Relaxed molecular clocks, the bias–variance trade-off, and the quality of phylogenetic inference. Syst. Biol. 59:1–8.
Whelan S., Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18:691–699.
Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.
Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to Primate Lysozyme evolution. Mol. Biol. Evol. 15: 568–573.
Yang Z., editor. 2006. Computational molecular evolution. New York: Oxford University Press.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.
Yang Z., Kumar S., Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141:1641–1650.
Yang Z., Roberts D. 1995. On the use of nucleic acid sequences to infer early branchings in the Tree of Life. Mol. Biol. Evol. 12: 451–458.
Zeldovich K.B., Berezovsky I.N., Shakhnovich E.I. 2007. Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput. Biol. 3:e5.
Zou L., Susko E., Field C., Roger A.J. 2012. Fitting nonstationary general-time-reversible models to obtain edge-lengths and frequencies for the Barry–Hartigan model. Syst. Biol. 61: 927–940.

## 2.2 Les modèles ECG (Empirical CAT-GTR) : efficacité de la modélisation de la variation du processus évolutif entre sites.

### 2.2.1 Introduction

Les modèles C10 à C60 empiriques publiés par Le et al. (2008a) ont pour but de prendre en compte l'hétérogénéité du processus évolutif entre sites tout en considérant un jeu de profils restreint en comparaison avec le modèle CAT (Lartillot and Philippe, 2004), qui a la liberté d'optimiser librement le nombre de profils et qui a tendance à en utiliser un très grand nombre. Avec les modèles C10 à C60, ce nombre de profils n'est pas optimisé et est fixé *a-priori* à 10, 20, 30, 40, 50 ou 60. Bien que le nombre limité de profils soit très probablement une sous-estimation du nombre optimal de profil, ces modèles peuvent être utilisés en tant que modèle de mélange dans le cadre du Maximum de Vraisemblance, du fait du nombre restreint de profils. A la suite de la publication de ces modèles, Le et al. (2008b) ont montré que les modèles C10 à C60 avaient en général des perfomances moindres quant à l'ajustement aux données par rapport aux modèles de mélanges de matrice de type UL3, alors que ces modèles ont un nombre de catégories dans le mélange bien inférieur (2 ou 3) (sauf dans le cas de données contenant beaucoup de saturation, où les modèles C10 à C60 retrouvaient des performances équivalentes aux modèles de mélange de matrices). Plusieurs raisons peuvent expliquer cela. Premièrement, lors de l'apprentissage des paramètres des profils à partir de la base de données HSSP (Schneider et al., 1997), Le et al. (2008a) ont rencontré des problèmes de répétabilité ; lorsque plusieurs optimisations étaient lancées en parallèle, elles ne convergeaient pas toutes vers les mêmes patrons de profils Le et al. (2008a). Il est possible que pour plusieurs jeux de données, certains des profils du mélange ne soient pas du tout adaptés en raison du fait qu'ils ne représentent pas un maximum global, et du coup entraine une diminution du fit aux données. Deuxièmement, et surement de manière plus importante, l'hypothèse d'échangeabilités constantes quelque soit le couple d'acides aminés considéré peut dans de nombreux cas s'avérer irréaliste. Par conséquent, nous avons voulu améliorer les modèles de mélanges de profils empiriques en suivant l'état d'esprit du modèle CAT-GTR, qui considère une matrice d'échangeabilités de type GTR au lieu d'une matrice non-informative de type Poisson (F81).

Les premiers tests ont consisté en l'utilisation de la matrice empirique d'échangeabilités LG, que l'on considérait commune à un ensemble finit de profils (de 10 à 60), comme dans les modèles de Le et al. (2008a). J'ai ensuite proposé de construire les profils directement à partir du jeu de données analysé, sans passer par une étape d'optimisation au Maximum de Vraisemblance. L'idée est de calculer les profils du modèles à partir des profils observés à chaque site. Beaucoup d'utilisateurs fixent les fréquences d'équilibres d'un modèle homogène de type

WAG (Whelan and Goldman, 2001) ou LG (Le and Gascuel, 2008) au fréquences observées pour les ré-ajuster par rapport au jeu de données. L'état d'esprit de ce que je proposais était le même. À partir des N profils observés, j'ai utilisé plusieurs approches de clustering afin de calculer un nombre limité de profils (K) représentatifs de tous les profils spécifiques de chaque site de l'alignement. Ces approches de clustering ont été codées dans Bio++ (excepté le clustering hiérarchique, déjà présent dans les librairies) afin d'utiliser directement les profils résultant du clustering comme profils du modèle de mélange. Dans le cas d'un jeu de données protéique de grande taille, comme dans le cas de concaténats, il est possible de construire ces profils à partir du jeu de données, car suffisamment d'information est présente pour apprendre les profils. Concernant les jeux de données plus courts, comme dans le cas d'alignement de gène unique, il était envisager d'utiliser la même approche à l'échelle de la base de données HSSP, et de fixer ensuite définitivement les profils afin d'être par la suite ré-utilisés. La figure 2.1 résume les différentes approches de clustering testées.

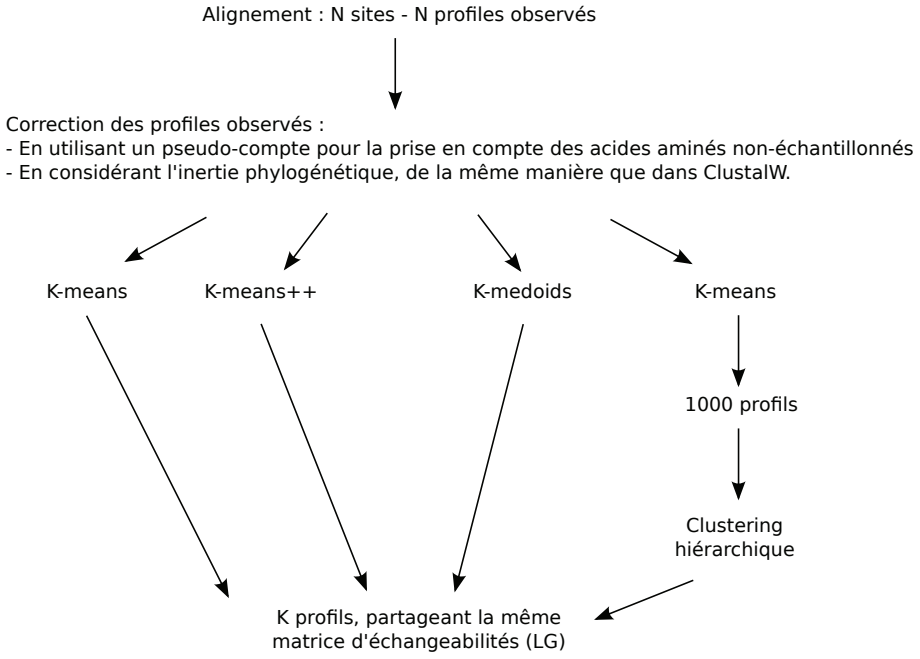



FIGURE 2.1 – Différentes procédures de clustering de profiles observés.

Le problème majeur que nous avons rencontré est également le manque de répétabilité dans le calcul des profils. En effet, les approches de clustering envisagées (K-means, K-means++, K-medoids) utilisent toutes K points de départ aléatoires dans l'espace des N profils. Il a été observé que le choix de ces points aléatoires influence grandement à la fois la patron des profils mais aussi la vraisemblance calculées avec le modèle de mélange utilisant ces profils. Même

l'algorithme K-means++, connu pour être moins sensible au point de départ, entraînait le même problème. En outre, la majorité des profils reconstruits étaient pour la plupart concentrés sur un seul acide-aminé et étaient donc non-informatifs d'un point de vue de la reconstruction phylogénétique, bien qu'ils permettaient d'atteindre d'excellentes vraisemblances. De même, plusieurs approches ont été testées afin de rendre les profils spécifiques d'un sous ensemble d'acides aminés (2 ou 3, comme observé dans le cas des modèles CAT ou C10 à C60). Par exemple, les sites entièrement ou très conservés étaient exclus du clustering. Malgré une amélioration visible, les profils observés n'étaient pas vraiment satisfaisant.

Cette approche aurait permis de calculer très rapidement des profils très adaptés au jeu de données. Malgré cela, cette idée a été abandonnée au profit d'une ré-estimation de modèles de mélange de profils de type CAT-GTR empiriques par un algorithme d'Expectation-Maximization (EM). Le manuscrit suivant présente la construction de ces modèles ainsi que de très bons résultats, permettant d'envisager une soumission très rapidement. Cependant, certains calculs sont actuellement toujours en cours, dont les résultats permettront la comparaison complète de tous les modèles concernés.

### 2.2.2 Manuscrit

# Efficient modeling of protein site-heterogeneities with empirical mixtures of profiles.

Mathieu Groussin[1], Laurent Guéguen[1], Bastien Boussau[1,2],
Manolo Gouy[1], and Nicolas Lartillot[1,3]

1 : *Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, Université Lyon 1, CNRS, UMR5558, Villeurbanne, France*

2 : *Department of Integrative Biology, University of California, Berkeley, United States of America*

3 : *Département de Biochimie, Université de Montréal, Montréal, Quebec, Canada*

## Abstract

Substitution models based on biological characteristics of sequence evolution capture more efficiently the complexity of evolutionary processes and produce more accurate phylogenetic histories. Site-specific selective constraints acting on protein sequences are key features of the evolutionary process, such that substitution models assuming heterogeneous processes along the sequence by considering several Markovian processes outperform site-homogeneous models that consider a single Markovian process for all sites. Such site-heterogeneous models are usually mixture models, that are either mixtures of Markovian matrices or mixtures of stationary probability profile. Recent years have seen a rapid expansion of studies showing that site-heterogeneous models of the kind of mixtures of profiles were able to solve difficult phylogenetic questions, especially by efficiently accommodating sequence saturation. Most of these studies used the site-heterogeneous CAT or CAT-GTR models in the Bayesian context. Empirical versions of these mixtures of profiles were previously described. Although they were proved to perform well on saturated datasets, their overall goodness of fit was weak in comparison with other types of site-heterogeneous models such as mixtures of matrices. This emphasizes the need to develop improved empirical mixtures of profiles for ML, in order to be applied on short alignments or large alignments in cases where Bayesian models fail to converge. Here, we extend the previously described empirical mixtures of profiles (models C10 to C60), which were empirical versions of the CAT model, by introducing an empirical GTR exchangeability matrix that is common to all profiles. In a way, the empirical model we propose mimic the CAT-GTR model in its spirit of assuming a constant short-term evolutionary process among sites while modeling the site-specific selective constraints with a fixed number of profiles. All evolutionay parameters were learned by an expectation-maximization algorithm on a dataset extracted from the HSSP database. We present seven empirical mixtures of profiles, which are different regarding their (fixed) number of profiles. We show that our ECG (Empirical CAT-GTR) models outperform any site-homogeneous or site-heterogeneous substitution models with respect to the fit of the data, both on short and large alignments. We used posterior predictive experiments to demonstrate that ECG models are more able to accommodate multiple substitutions, making them potentially more robust to phylogenetic artifacts. ECG models are implemented in both Bayesian and ML contexts and are available in Phylobayes and the Bio++ libraries, respectively.

## Introduction

Markovian substitution models are employed to describe biological sequence evolution and generally have parameters that aim at phenomenologically describing the different rates of change between character states. They have been used in a variety of fields such as, to name a few, the reconstruction of phylogenies (Yang, 2006), the dating of species divergences (Thorne et al., 1998; Drummond et al., 2006; Yang and Rannala, 2006), the detection of natural selection (Nielsen and Yang, 1998), the inference of ancestral population sizes (Dutheil et al., 2012) or the inference of ancestral molecular compositions or ancestral sequences (Gaucher et al., 2003; Boussau et al., 2008; Harms and Thornton, 2010; Groussin and Gouy, 2011). The parameters of the model may be estimated for each dataset analyzed, as it is usually the case with DNA or codon models or once from a large dataset and then fixed to be subsequently re-used on other datasets. Depending on how parameters are treated, models are referred as mechanistic or empirical, respectively. Time-reversible Markovian models are characterized by a rate matrix providing instantaneous substitution rates between amino acids. This rate matrix is the product between the matrix of relative exchange rates (or exchangeabilities) and the diagonal matrix of equilibrium frequencies often named profile.

In most cases, it is assumed that the evolutionary process is homogeneous among sites, and the same substitution model is used to describe the substitution history for every site. However, when looking at a protein alignment, one can immediately observe that sites do not share the same sets of amino-acids. Although there are 20 possible amino acids, a particular site is usually characterized by only a small subset of these amino acids, varying all along the sequence (Lartillot and Philippe, 2004). Site-specificities can result from biochemical or structural constraints, which restrict the possible amino acids retained by selection. Consequently, as the pattern of amino acid is heterogeneous among sites, the evolutionary process varies accordingly. The use of site-homogeneous models containing a single matrix of substitution rates such as JTT (Jones et al., 1992), WAG (Whelan and Goldman, 2001) or LG (Le and Gascuel, 2008) can then become problematic due to their unrealistic assumption of a constant process acting on all sites of a protein. For instance, it has been shown that site-homogeneous models have a higher susceptibility to be affected by the Long-Branch Attraction artefact than site-heterogeneous models which allow the process to vary between the columns of an alignment (Lartillot et al., 2007). As sites usually undergo successive substitutions among a small subset

of the 20 amino acids, homoplasies between these site-specific amino acids will be high. By modeling site evolution with a set of different profiles which may be more specific of a small number of amino acids than profiles of single matrix models, site-heterogeneous models thus estimate more properly the level of homoplasy and interpret less frequently convergences as shared ancestry.

It has been proposed to use mixture models containing several Markovian models (or components) to deal with the variation of the evolutionary process among sites (Koshi and Goldstein, 1998; Pagel and Meade, 2004; Gascuel and Guindon, 2007). These mixture models can have different properties. (i) The mixture model may have a fixed or an infinite dimension, depending on whether the number of components is set *a priori* or is a parameter of the model. (ii) The assignment of a given site may be to a particular component or probabilistic on all components *a priori*. In the latter case, the likelihood of the site is the sum of the weighted likelihoods computed with each component and the site may have *a posteriori* probabilities to belong to a given component. (iii) The mixture model may be empirical, meaning that all entries of the substitution matrices present in the mixture are pre-estimated on a large dataset and subsequently re-used on other datasets. Empirical models are a good solution to avoid overfitting issues when the size of the dataset is too short to accurately estimate all parameters of the model. Alternatively, the mixture model may be mechanistic. In this case, all parameters of the different models are directly estimated from the dataset under study. If this dataset is large enough, the advantage is that the model will more efficiently fit the data and potentially lead to more accurate phylogenetic estimations.

Mechanistic (Koshi and Goldstein, 1998) or empirical (Thorne et al., 1996; Goldman et al., 1998; Liò and Goldman, 1999) mixture models with an a priori fixed number of categories have been proposed. They take into account protein properties that are heterogeneous along the sequence and that influence the substitution process such as solvent exposure or secondary structure. In line with this, Le et al. (2008b) and Le and Gascuel (2010) proposed a series of empirical mixture models with fixed dimensionality that outperform any single matrix models. These models were learned on the HSSP database (Schneider et al., 1997) by taking into account variation of rates among sites, in a supervised or unsupervised way. In the supervised way, sites sharing solvent exposure or secondary structure properties were *a priori* assigned to specific components. Four models were inferred in this way:

- EX2, which is composed of two matrices corresponding to exposed/buried sites
- EX3, which is composed of three matrices corresponding to highly exposed/intermediate/buried sites
- EHO, which is composed of three matrices corresponding to extended/helix/other sites
- EX_EHO, which is composed of six matrices corresponding to the combination of EX2 and EHO.

In the unsupervised way, both site partitions and their corresponding matrices were directly learned from the data. Two models were proposed:

- UL2, which is composed of two matrices
- UL3, which is composed of three matrices

Note that all these models are mixtures of matrices with both exchangeabilities and equilibrium frequencies varying among components.

Lartillot and Philippe (2004) proposed a Bayesian site-heterogeneous mixture model with an infinite dimensionality and in which a site is allocated to a given category of the mixture. This mechanistic model, named CAT, makes the assumption that the heterogeneity lies entirely in equilibrium frequencies and that all categories of the mixture share a common Poisson (or F81) exchangeability matrix, where all exchangeabilities are equal. CAT is said to be a mixture of profiles, as only equilibrium frequencies differ among sites. The pattern of the different profiles is directly learned from the data through a Dirichlet process. The CAT model was proved to improve phylogenetic inferences, notably due to its lower sensitivity to LBA. CAT was then extended by considering a GTR exchangeability matrix instead of the Poisson process, with exchangeabilities directly estimated from the data. However, the CAT-GTR model may suffer from two major issues. First, a certain amount of data is necessary to accurately estimate all free parameters. Second, problems of convergence may appear when the dataset is too large, like with protein concatenates.

To circumvent these problems, six empirical versions of the CAT model with finite dimensions (from 10 to 60) have been proposed (Le et al., 2008a). These mixtures of profiles were estimated on the HSSP database through an expectation-maximization (EM) algorithm. Although it was shown that 20 profiles were sufficient to better fit the data than single matrix

models, the C10 to C60 models tend to be outperformed by the mixtures of matrices proposed by (Le et al., 2008b), underlining the need for the model to account for the short-term substitution processes acting on amino-acids with exchangeabilities.

In this paper, we propose a new set of 7 empirical mixtures of profiles learned on the HSSP database with a modified version of the EM algorithm described in Le et al. (2008a), which was adapted to consider a GTR exchangeability matrix instead of a Poisson process. Each model is an empirical version of the Bayesian CAT-GTR model which accounts for the variation of evolutionary processes across sites through a set of different profiles and which suppose that all sites share similar exchangeabilities. The models contain either 6, 10, 20, 30, 40, 50 or 60 profiles. These models are hereafter named ECGX for Empirical CAT-GTR, the letter X standing for the number of profiles. We show that ECG models tend to yield highly significant likelihood gains and globally outperform any other single-matrix or mixture models, with both single protein or phylogenomic datasets.

## Material and Methods

### Notation, Data and Model

The data consist of $P = 1015$ single-gene sequence alignments, taken from the HSSP database. Each alignment has its own set of taxa. For each alignment, the tree topology, the branch lengths and the shape parameter of the distribution of rates of substitution across sites are first separately estimated by maximum likelihood under the LG model (using phyml), and then considered as fixed for the rest of the calculation. In the following, genes are ordered and amino-acid positions are globally indexed by $i = 1..N$, where $N$ is the sum of the lengths of all $P$ alignments. The tree topology, the branch lengths and the shape parameter of the distribution of rates of substitution across sites of the gene containing position $i$, such as optimized by maximum likelihood under the LG model, are collectively denoted as $\hat{\lambda}_i$.

A mixture model is defined, with a fixed number $K$ of components. Each component $k = 1..K$ defines a general time-reversible amino-acid replacement process, characterized by its $S \times S$ rate matrix $Q_k$ and its own set of equilibrium frequencies $\pi_k = (\pi_{ka})_{a=1..S}$, where $S = 20$ is the number of states of the substitution process. All components share the same set of relative exchange rates $\rho = (\rho_{ab})$, such that $\rho_{ab} = \rho_{ba}$. The amino-acid replacement matrix

of component $k$ is therefore:

$$\begin{aligned} Q_{kab} &= \rho_{ab}\pi_{kb}, \\ Q_{kaa} &= -\sum_{b\neq a} Q_{ab}. \end{aligned}$$

Component $k$ has prior weight $w_k$, $k = 1..K$, such that $\sum_k w_k = 1$. The free parameters of the model are therefore $w = (w_k)_{k=1..K}$, $\rho = (\rho_{ab})_{1\leq a<b\leq S}$ and $\pi = (\pi_k)_{k=1..K}$, where for each $k$, $\pi_k = (\pi_{ka})_{a=1..S}$. Collectively, all these parameters are denoted as $\theta = (w, \rho, \pi)$.

The overall likelihood function is therefore a product over all sites, in which the likelihood at each site is a weighted sum over all components of the mixture:

$$\begin{aligned} L(\theta) &= \prod_i p(C_i \mid \hat{\lambda}_i, \theta) \\ &= \prod_i \sum_k w_k p(C_i \mid \hat{\lambda}_i, \rho, \pi_k). \end{aligned}$$

Each site- and component-specific likelihood factor is in turn a sum over all possible substitution histories at site $i$:

$$p(C_i \mid \hat{\lambda}_i, \rho, \pi_k) = \sum_{\Xi_i | C_i} p(\Xi_i \mid \hat{\lambda}_i, \rho, \pi_k).$$

In this equation, the sum is over all substitution histories $\Xi_i$ compatible with column pattern $C_i$ ($\Xi_i \mid C_i$). In practice, this sum has both discrete and continuous aspects (that is, it is a sum of integrals). For simplicity, we denote it as if it was purely discrete.

### Expectation-Maximization

The aim is to maximize $L(\theta)$ with respect to $\theta$. To do this, we use an approximate EM procedure. As in Le et al. (2008a), the expectation of the logarithm of the likelihood is taken simultaneously over the allocations of sites to the components of the mixture and over the substitution histories at each site. Denoting by $z = (z_i)_{i=1..N}$ the allocation vector (such that, for all $i$, $z_i \in [1..K]$ is the component to which site $i$ is allocated), the expected log likelihood is:

$$E_{z,\Xi}[\ln p(\Xi \mid \hat{\lambda}, z, \rho, \pi)]. \tag{1}$$

In the exact EM approach, equation 1 is an expectation over the posterior distribution for $z$ and $\Xi$ conditional on the current parameter value $\theta^*$, which is then maximized with respect to $\theta$. The procedure is iterated until numerical stabilization, at which point a (potentially local) maximum of the likelihood function has been reached.

Conditional on the current allocation, and for a Markov substitution process, the expectation of the logarithm of the augmented likelihood over all substitution histories at a given site can be expressed as a function of a few fundamental statistics: probability of being in each possible state at the root, expected total time spent in each possible state, and expected number of transitions between each pair of states. F81 processes allow for algorithmic shortcuts leading to fast computation of these expected elementary statistics over substitution histories (Le et al., 2008a). In the general time-reversible case, these expectations are analytical (Holmes and Rubin, 2002). However, their computation is time-consuming. Specifically, whereas the classical pruning algorithm used for integrating the likelihood over all substitution histories is linear in the number of states $S$ (here $S = 20$ amino-acids), computing the expectations over substitution histories is quadratic in $S$, thus effectively representing a 400-fold increase in computational complexity, compared to plain likelihood evaluation.

This computational bottleneck motivated the following approximation: first, for each gene of the training set, the expected statistics over the substitution histories are computed separately for each site under the LG model and under the maximum likelihood parameter estimates $\hat{\lambda}$, using essentially the method described in Holmes and Rubin (2002). These expectations are then considered as fixed, and the EM is conducted only by recomputing the expectation over allocations at each iteration. In this way, the limiting step represented by the computation of the elementary expectations over substitution histories is done only once. This approach is expected to result in reasonable estimates as long as substitution histories are globally robust to the choice of the underlying model and parameter values.

Mathematically, we define the following statistics:

- $q_{ia}$: the probability for site $i$ to be be in state $a$ at the root;
- $t_{ia}$: the expected total time site $i$ has been in state $a$ along the tree (note that this time is averaged over the 4 categories of the discretized gamma distribution; for each category, the time is scaled by the corresponding relative substitution rate);

- $n_{iab}$: the expected number of substitutions between states $a$ and $b$ at site $i$, along the tree.

The expected log-augmented likelihood at site $i$, assuming this site to be allocated to component $k$ (characterized by the matrix $Q_k$ and equilbrium frequency profile $\pi_k$) is then equal to:

$$\langle \ln L_{ik} \rangle \quad = \quad \sum_a q_{ia} \ln \pi_{ka} \; - \; \sum_a t_{ia} |Q_{kaa}| + \sum_{ab} n_{iab} \ln Q_{kab}.$$

This suggests to define the pseudo-likelihood at site $i$, conditional on allocation to component $k$, as:

$$\tilde{L}_{ik} \quad = \quad e^{\langle \ln L_{ik} \rangle}.$$

This likelihood can then be averaged over all components for site $i$

$$\tilde{L}_i \quad = \quad \sum_k w_k \tilde{L}_{ik}$$

and multiplied over all sites:

$$\tilde{L} \quad = \quad \prod_i \tilde{L}_i.$$

Maximizing this pseudo-likelihood with respect to the parameters of the mixture ($\rho$, $\pi$ and $w$) is done by EM, although now recomputing at each cycle the expectation only with respect to the allocation vector $z$.

The posterior allocation probabilities are given by:

$$p_{ik} \quad = \quad \frac{w_k \tilde{L}_{ik}}{\sum_l w_l \tilde{L}_{il}},$$

so that the expectation over allocations is:

$$\begin{aligned} E_z[\ln \tilde{L}] \quad &= \quad \sum_i \sum_k p_{ik} \langle \ln L_{ik} \rangle \\ &= \quad \sum_i \sum_k p_{ik} \left[ \sum_a q_{ia} \ln \pi_{ka} \; - \; \sum_a t_{ia} |Q_{kaa}| + \sum_{ab} n_{iab} \ln Q_{kab} \right] \\ &= \quad \sum_i \sum_k p_{ik} \left[ \sum_a q_{ia} \ln \pi_{ka} \; - \; \sum_a t_{ia} \sum_{b \neq a} \rho_{ab} \pi_{kb} + \sum_{ab} n_{iab} (\ln \rho_{ab} + \ln \pi_{kb}) \right]. \end{aligned}$$

The previous equation is separately maximized with respect to each of the subcomponents of $\theta$. That is, at each cycle, maximization is done either with respect to $\pi$, or to $\rho$, or to $w$, cycling over the three components until numerical stabilization.

Gathering terms in $\pi_k$ and defining:

$$\begin{aligned} U_{ka} &= \sum_i p_{ik} q_{ia} + \sum_i p_{ik} \sum_{b \neq a} n_{iab}, \\ B_{ka} &= \sum_i p_{ik} \sum_{b \neq a} (t_{ia} + t_{ib}) \rho_{ab}, \end{aligned}$$

leads to:

$$E_z[\ln \tilde{L}] = \sum_k \left[ \sum_a U_{ka} \ln \pi_{ka} - \sum_a B_{ka} \pi_{ka} \right] + \ldots,$$

where terms not depending on $\pi$ have been dropped. Each $\pi_k$ can therefore be optimized by maximizing $\sum_a U_{ka} \ln \pi_{ka} - \sum_a B_{ka} \pi_{ka}$, under the constraint that $\sum_a \pi_{ka} = 1$ and $\pi_{ka} > 0$ for all $a$. To do this, the following Lagrangian is defined:

$$\mathcal{L} = \sum_a U_{ka} \ln \pi_{ka} - \sum_a B_{ka} \pi_{ka} - \beta \left( \sum_a \pi_{ka} - 1 \right),$$

and is differentiated with respect to $\pi_{ka}$ and $\beta$:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \pi_{ka}} &= \frac{U_{ka}}{\pi_{ka}} - B_{ka} - \beta, \quad a = 1..S, \\ \frac{\partial \mathcal{L}}{\partial \beta} &= 1 - \sum_a \pi_a. \end{aligned}$$

The derivative with respect to $\pi_{ka}$ has solution $\pi_{ka} = U_{ka}/(B_{ka} + \beta)$. Substituting this into $\frac{\partial \mathcal{L}}{\partial \beta}$ gives:

$$\mathcal{D} = \frac{\partial \mathcal{L}}{\partial \beta} = 1 - \sum_a \frac{U_{ka}}{B_{ka} + \beta}.$$

The root of $\mathcal{D}$ is found numerically using the Newton-Raphson procedure. Specifically, noting that

$$\mathcal{D}' = \frac{\partial^2 \mathcal{L}}{\partial \beta^2} = \sum_a \frac{U_{ka}}{(B_{ka} + \beta)^2},$$

we define the iteration:

$$\beta_{n+1} = \beta_n - \frac{\mathcal{D}}{\mathcal{D}'},$$

which quickly converges (within 10 to 20 steps) to a numerical estimate of the root $\hat{\beta}$ of $\mathcal{D}$. Finally, the solution of the constrained maximization is obtained by substituting this root into the expression for $\pi_k$:

$$\pi_{ka} = U_{ka}/(B_{ka} + \hat{\beta}).$$

To optimize $\rho$, terms in expression of $E_z[\ln \tilde{L}]$ have to be gathered differently, by defining:

$$\begin{aligned} V_{ab} &= \sum_i n_{iab} + n_{iba}, \\ G_{ab} &= \sum_i \sum_k p_{ik}(t_{ia}\pi_{kb} + t_{ib}\pi_{ka}), \end{aligned}$$

such that

$$E_z[\ln \tilde{L}] = \sum_{a<b} V_{ab} \ln \rho_{ab} - \sum_{a<b} G_{ab}\rho_{ab} + \ldots,$$

where terms not depending on $\rho$ have been omitted. This can be maximized, with respect to $\rho$, simply by setting $\rho_{ab} = G_{ab}/V_{ab}$.

Finally, optimizing the weights of the mixture $w$ is done by setting:

$$w_k = \frac{\sum_i p_{ik}}{\sum_k \sum_i p_{ik}}.$$

In summary, the overall method therefore consists of the following steps:

- calculating the posterior allocation probabilities $p_{ik}$ based on the pseudo-likelihood (eq 2),
- optimize either one of $\pi$, $\rho$, or $w$,
- iterate until the absolute difference between successive scores is less than 0.1 (for a total final log likelihood over a set of 1015 alignments of the order of $\ln L = -6.10^6$).

### Phylogenomics datasets

ECG models have been run on four phylogenomic datasets to test their ability to efficiently fit the data on large concatenates in comparison with other site-homogeneous and site-heterogeneous empirical models. These datasets comprise:

- a universal dataset used in Boussau et al. (2008), containing 3,336 sites and was obtained by concatenating 56 gene families spanning 30 representative species of the Tree of Life. The topoloy in Figure 2 of Boussau et al. (2008) was considered.
- an archaeal dataset, which is a concatenation of 72 protein-coding genes sampled in 35 archaeal species and 10 bacterial species (Groussin and Gouy, 2011). We removed Bacteria from the alignment, as well as the two uncultured thaumarchaeal species, for which only one protein sequence was present in the alignment. The final 9,387 amino acid-long dataset contains 33 archaeal species. We used the topology in Figure 3 of Groussin and Gouy (2011) to run ECG models.
- a eukaryotic dataset, composed of 129 proteins sampled in 36 species (Douzery et al., 2004). The alignment contains 30,399 positions. The topology in Figure 1 of Douzery et al. (2004) was used.
- a metazoa dataset, containing 128 concatenated proteins with 30,257 positions (Philippe et al., 2009). The topology in Figure 1 of Philippe et al. (2009) was used.

### Model selection

We used the AIC (Akaike, 1973) and the second order AIC (AICc) (Sugiura, 1978) model selection criteria to select the best-fitting model among non-nested models. These criteria are computed as follows:

$$AIC = -2 \times \ln L + 2 \times p$$

$$AICc = AIC + \frac{2 \times p \times (p+1)}{I - p - 1}$$

$$BIC = -2 \times \ln L + p \times \ln I$$

with $p$ the number of parameters estimated by ML and $I$ the total number of sites per alignment. The use of the second order AIC criterion (AICc) is recommended with respect to

AIC when the size of the alignment is small, as it may happen with HSSP single gene datasets (Burnham and Anderson, 2004). AICc applies a greater penalty for extra parameters to avoid the selection of models containing too many parameters with small alignments. Note that when $I >> p$, AICc converges to AIC. BIC penalizes parameter-rich models far more severely than does AIC or AICc and was suggested to penalize too strongly empirical mixtures of profiles (Le et al., 2008a). Consquently, BIC was only applied on phylogenomic datasets.

All ML calculations were performed with a discrete $\Gamma$ distribution and 4 categories to model the variation of evolutionary rates among sites. With a single empirical matrix model (LG), the $\alpha$ parameter of the $\Gamma$ distribution is the only parameter ($p = 1$). All mixture models were run with the ML optimization of the weight of each component. Consequently, for a given mixture model with $K$ components, $K - 1$ additional parameters are accounted for in the calculation of $p$.

### Availability

ECG models have been implemented in both Bayesian and ML contexts. They are available in Phylobayes (Lartillot et al., 2009) and in the Bio++ libraries (Guéguen et al., 2013) to be used with the bppML program (Dutheil and Boussau, 2008).

## Results

### Characteristics of the ECG models

Interestingly, the exchangeabilities of the different ECG models strongly correlate between each other (all correlation coefficients are equal to $r^2 = 0.99$). Besides, the ECG exchangeabilities also strongly correlate to those of the LG model (again, all correlation coefficients are equal to $r^2 = 0.99$), which was inferred on the Pfam database (Bateman et al., 2002). It indicates that the LG and ECG exchangeability matrices describe similar short-term substitution processes, where amino acids sharing similar biological, chemical, and physical properties tend to exchange more frequently between each other. Le and Gascuel (2008) observed that LG and WAG exchangeabilities highly correlate as well. However, the relative difference between amino acids revealed that the two matrices were actually quite different, with some exchangeabilities

of one of the two matrices being up to 6 times higher than in the other matrix. Accordingly, we observed that the mean of absolute relative differences between WAG and LG exchangeabilities (a relative difference being defined as $(\mathrm{LG}_{ij} - \mathrm{WAG}_{ij})/(\mathrm{LG}_{ij} + \mathrm{WAG}_{ij})$, with $i$ and $j$ two amino acids) is $0.21+/-0.16$. In our case, ECG exchangeabilities are generally much closer to those of LG: the mean of absolute relative differences is $0.06+/-0.06$ for all ECG models. One of the differences between WAG and LG is that variation of rates among sites is accounted for by LG. It was also taken into account in our EM algorithm. Therefore, it shows that the presence of several profiles (ECG) instead of only one (LG) does not influence the description of exchangeabilities, whose signal seems to be robust. More importantly, it highlights that the good performances in terms of data fitting for ECG models in comparison with LG is for the most part due to the presence of profiles, which adds a considerable amount of information and allows to better discriminate homoplasies from true phylogenetic signal.

Le et al. (2008a) showed that the profiles of the C10 to C60 models are generally specific to two or three amino acids that have high equilibrium frequencies and that tend to have similar biological characteristics. However, in our case, the profiles of the different ECG models do not have this property. They are generally specific to more than 3 amino acids that are not necessarily biologically similar. It can be easily explained by the presence of the exchangeability matrix, which is capable of capturing the short-term substitutional effect making similar amino acids more exchangeable. Thus, ECG models can more freely accept that amino acids having different properties co-exist within a single profile, because these amino acids will less likely substitute between each other owing to their low exchangeabilities. For instance, one of the profiles of the ECG20 model gives high frequencies for A, D, E and K. D and E are both negatively-charged, K is positively-charged and A is hydrophobic. In line with this, exchangeabilities between biochemically distinct amino acids are low ($\rho_{A\leftrightarrow E} = 1.18, \rho_{A\leftrightarrow D} = 0.5, \rho_{A\leftrightarrow K} = 0.59, \rho_{K\leftrightarrow D} = 0.26, \rho_{K\leftrightarrow E} = 1.58$) but the exchangeability between D and E is higher ($\rho_{D\leftrightarrow E} = 5.18$). One of the reason that may explain why C10 to C60 models have moderate performances in terms of fit to the data, is that their constraint on a fixed number of profiles prevent them to capture a large part of the complexity of the evolutionary process among sites. Regarding ECG models, which have the same constraint, they outperform C10 to C60 models (see below) not only because they give information on the exchangeabilities by themselves, but also because these exchangeabilities allow the model to capture more efficiently the site-wise heterogeneities by concentrating several site patterns

within a single profile.

### ECG models are the most efficient at fitting the data

To evaluate the capacity of ECG models to efficiently fit the data, we run bppML (Dutheil and Boussau, 2008) on each of the 1,030 alignments of our restricted HSSP database (see Materials and Methods), with the ECG6, 10, 20, 30 and 60 models. ECG models were compared to the LG site-homogeneous models and to site-heterogeneous mixtures of matrices (EX2, EX3, EHO, UL2, UL3 and EX_EHO) and to previous mixtures of profiles (C10 and C20). For all mixture models, goodness of fit was measured with gains in AIC and AICc criteria with respect to LG. Figure 1 shows that all mixtures of matrices and ECG models fit better the data than LG. This highlights the need to account for the heterogeneity of the evolutionary process among sites. However, the C10 mixture of profiles tend to be, on average, outperformed by LG, supporting previous observations (Le et al., 2008b). ECG models yield better performances than any other mixtures in terms of average gain in AIC and AICc, with the exception of ECG6. In comparison with previous mixtures of profiles (C10-C60), these results strongly support the need to account for exchangeability information in substitution models. It further indicates that describing site heterogeneity with only 6 profiles is not efficient enough to capture all specificities of the different processes acting along the sequences.

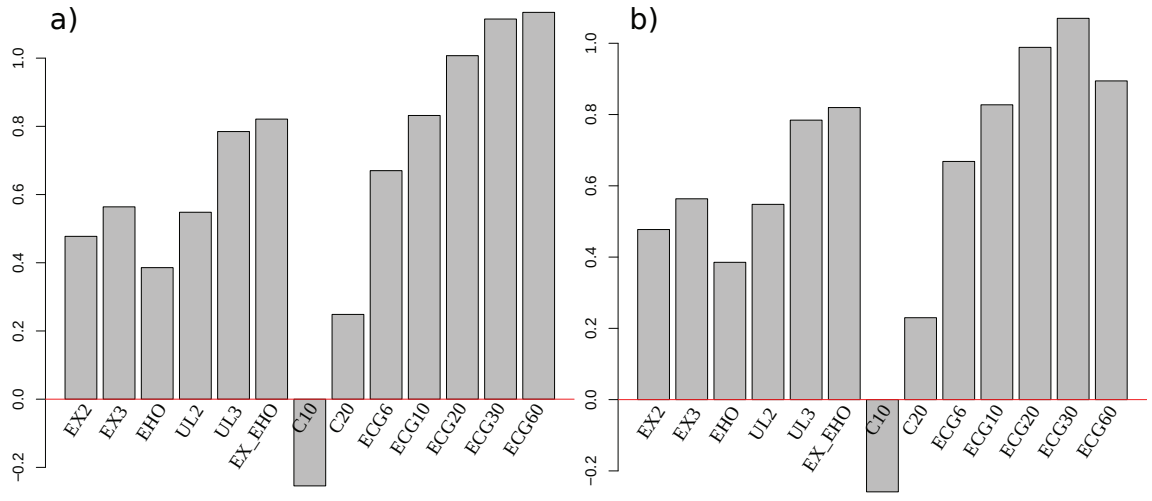



Figure 1: **Average gain in model fit to the data**. Average gains per site over 1,030 alignments in AIC (a) and AICc (b) are represented, with LG as a reference. The six first models are mixtures of matrices. The seven others are mixtures of profiles.

As previously observed (Le et al., 2008a), the major gains in fit occured when going from ECG6 to ECG30, suggesting that even with a relatively small number of profiles (10 to 30), a sufficient part of site heterogeneities can be captured. Nonetheless, these numbers are usually far from the numbers of profiles produced by the CAT or CAT-GTR models, which are free to estimate the optimal number of required profiles given the data under analysis.

We also assessed the performance of mixture models with the number of alignments for which they provide a better fit than LG. Figure 2 confirms previous observations concerning the poor capacity of C10 and C20 models to efficiently fit the data compared to ECG10 and ECG20. However, Figure 2 also highlights that there is a progressive decrease in the number of alignments on which ECG fits better than LG with higher number of profiles. When the number of profiles is higher than 20, LG provides a better fit than ECG models on several alignments.

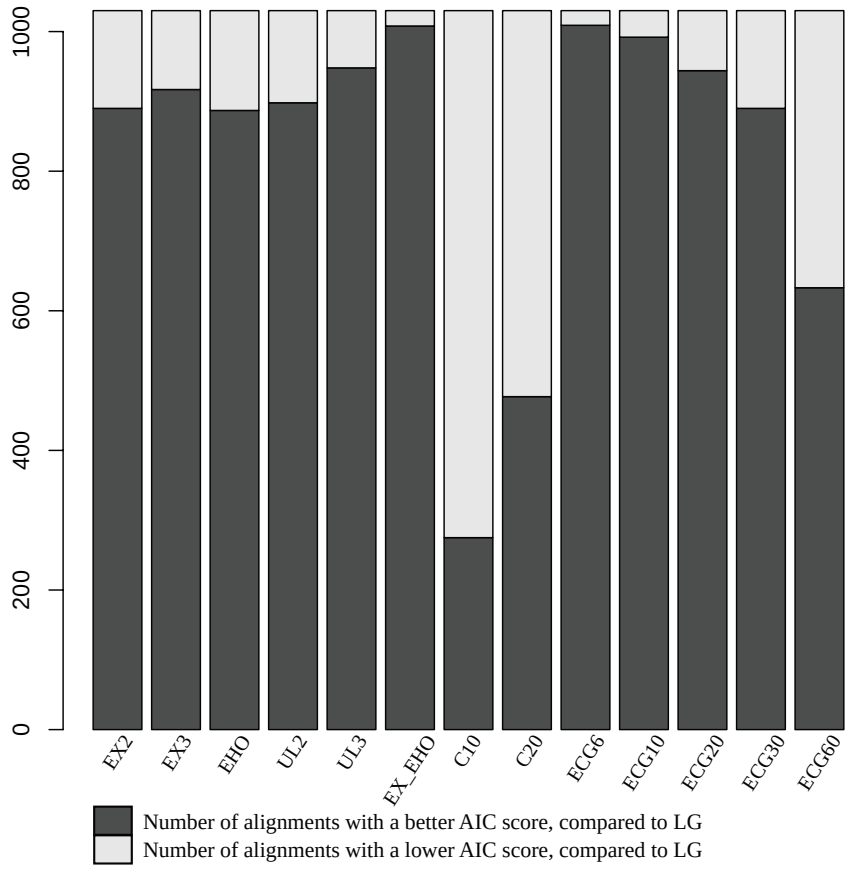



Figure 2: **Number of alignments for which mixture models are better than LG**. The results obtained with the AICc criterion are represented. Other criteria give similar tendencies.

The previous results presented in Figure 1 indicate that for cases where LG has a better fit than ECG, the loss of fit for ECG is small. On the contrary, when the ECG provides a gain of fit with respect to LG, this gain is very high. For instance, with the ECG30, the average gain in AICc is 317 points while the average loss is 23 points. As mixtures of profiles have been proved to be very efficient at coping with mutational saturation (Lartillot and Philippe, 2004; Le et al., 2008a), it suggests that some single gene alignments in our HSSP database may be characterized by a relatively slow rate of evolution, such that the high number of parameters in ECG models is much more penalized by model selection criteria. To verify this, we divided the 1,030 HSSP alignments into four sets of alignments depending on saturation degree. We used the total length of a tree to approximate saturation, assuming that alignments that produce long trees are more affected by mutational saturation. Figure 3 clearly shows that with the slowest alignments, mixture models tend to perform poorly in comparison with LG, with ECG30 and ECG60 even worse than other mixtures.

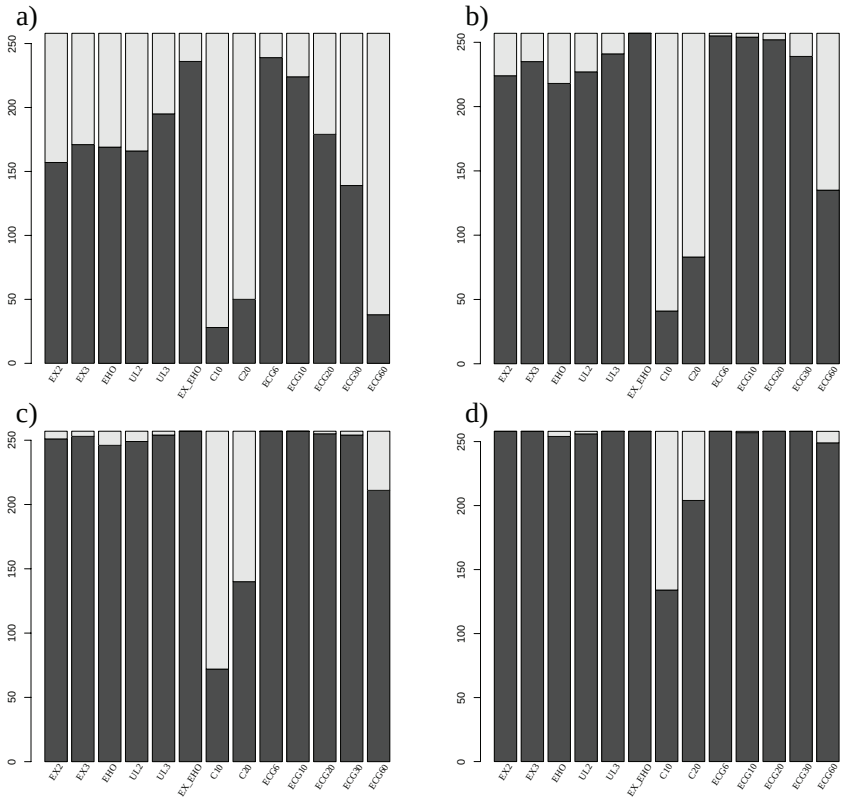



Figure 3: **Model fit performance depends on saturation**. For a given alignment, the saturation degree has been approximated with the total length of its corresponding tree computed with PhyML. HSSP alignments were classified into four quartiles, depending on the saturation degree. a): first quartile; b): second quartile; c): third quartile; d): fourth quartile. Colors are as in Figure 2.

However, when saturation increases, all mixture models (excepted C10 and C20) strongly outperfom LG. Regarding ECG60, a gradual increase is observed up to almost 100% of alignments where ECG60 is better than LG on the fourth quartile of saturation degree. This shows that the numerous profiles described in the model brings sufficient information to detect homoplasies owing to saturation, such that the increase in likelihood offsets the high number of parameters.

We used posterior predictive tests to verify if our ECG models accurately describe the evolutionary process, If so, the models should be able to reproduce characteristics of the real data. We compared four models implemented in Phylobayes: the site-homogeneous GTR model and three site-heterogeneous models, UL3, ECG30 and CAT-GTR. We focused on the ability of these models to reproduce the observed site-specific diversity, which is similar to the saturation index presented in (Lartillot et al., 2007). The diversity of a site is defined as the absolute number of different amino acids observed at this site. The rationale of using this statistic is that sites usually contain only four to five amino acids and that a model which is inadequate to describe these site-specific properties will tend to produce higher site-specific diversities than those observed on real data. Table 1 shows that the ECG30 model outperforms GTR and UL3 in its ability to adjust to the data. The average predictive site-specific diversity is closer to the observed diversity in data simulated with ECG30 ($p\text{-}value < 0.001$, student paired test between UL3 and ECG30), underscoring the higher ability of ECG30 to account for site-specific biochemical specificities. (Lartillot et al., 2007) showed that the capacity of the site-heterogeneous models to accurately adjust to these site-specificities explain why these models are more able to detect homoplasies. In line with this, the average tree length is higher with ECG30 than with GTR and UL3 ($p\text{-}value < 0.01$, student paired test between UL3 and ECG30), highlighting that ECG is more prone to efficiently detect hidden substitutions. However, it is worth noting that CAT-GTR strongly outperforms other empirical site-heterogeneous models with respect to these two criteria.

| | GTR | UL3 | ECG30 | CAT-GTR |
|---|---|---|---|---|
| Posterior predictive observed diversity | 1.1 | 0.79 | 0.7 | 0.16 |
| Tree Length | 14.3 | 14.2 | 15.2 | 18.9 |

Table 1: **Model comparison with respect to posterior predictive tests for site-specific diversity and tree lengths.** Figures represent average values for a sample of 100 out of the 1,030 alignments present in our HSSP database. For each model, the site-specific diversities represent the diffence between average posterior predictive site diversities produced by the model and the average observed site diversity.

Phylogenomic datasets, independent of the HSSP alignments, were also used to compare the mixture models. Table 2 illustrates that systematically, ECG models strongly outperform any other site-homogeneous or site-heterogeneous models, and that, among ECG models tested, ECG20 appears to be the best at fitting the data.

## Discussion

In this study, we described a set of empirical mixtures of profiles that aim to improve the modeling of site heterogeneities in comparison with previous mixture models. Importantly, we showed that our new ECG models greatly outperform the previously published mixtures of profiles (C10 and C20 in this paper) presented in Le et al. (2008a) and in which substitutions occur according to a Poisson (or F81) process. The Poisson process prevents to concretely discriminate the short-term selective constraints from the long-term constraints as it stipulates that all exchangeabilities are uniform and that all substitution processes are modeled in the profiles. With the CAT and C10 to C60 models, this approach appeared to be very efficient at correcting saturation due to site-specific constraints, as it is supposed that biological constraints on site-specific properties impact frequency patterns at a long evolutionary scale. However, estimating exchangeabilities with a GTR exchangeability matrix in combination with CAT was shown to be of great importance to increase the fit of the model to the data without reducing the capacities to accomodate saturation. The empirical versions of the CAT-GTR model presented in this study display high capacities of goodness of fit and are more able to efficiently account for saturation than other site-homogeneous or site-heterogeneous models.

The development of empirical mixture models faces the issue of how site-specific selective constraints should be modeled and how, conceptually, site heterogeneities should be differentially described between exchangeabilities and equilibrium frequencies. In the ECG models presented here, site-specific processes are only described by the profiles and all categories share the same exchangeabilities. In Markovian substitution models, exchangeabilities represent the general propensities of amino acids. They account for the short-term evolutionary processes acting on sequences and are more uniform along the sequence than the amino acid frequencies are. However, the mixture of matrices described by Le et al. (2008b), for which exchangeabilities differ among categories of the mixture, show that exchangeability heterogeneities are important. They showed that exchangeabilities among categories are much more correlated

| Dataset | model | log Likelihood | AIC score | BIC score |
|---|---|---|---|---|
| Universal | WAG | -146605.7 | 293213.5 | 293219.6 |
| | LG | -145652.8 | 291307.6 | 291313.8 |
| | EX2 | -144856.6 | 289717.2 | 289729.4 |
| | EX3 | -144672.5 | 289351.0 | 289369.4 |
| | EHO | -145057.2 | 290120.3 | 290138.7 |
| | UL2 | -144810.5 | 289624.9 | 289637.2 |
| | UL3 | -144372.4 | 288750.9 | 288769.2 |
| | EX_EHO | -144112.8 | 288237.6 | 288274.2 |
| | ECG6 | -144470.4 | 288952.7 | 288989.4 |
| | ECG10 | -144041.1 | 288102.2 | 288163.3 |
| | **ECG20** | **-143736.0** | **287512.1** | **287634.3** |
| Archaea | WAG | -343074.5 | 686151.0 | 686158.1 |
| | LG | -340369.2 | 680740.4 | 680747.5 |
| | EX2 | -338492.3 | 676988.6 | 677002.9 |
| | EX3 | -337948.5 | 675903.0 | 675924.5 |
| | EHO | -339144.1 | 678294.1 | 678315.6 |
| | UL2 | -337901.3 | 675806.6 | 675820.9 |
| | UL3 | -336378.7 | 672763.4 | 672784.8 |
| | EX_EHO | -336873.6 | 673759.2 | 673802 |
| | ECG6 | -337208.4 | 674428.7 | 674471.6 |
| | ECG10 | -335996.0 | 672011.9 | 672083.4 |
| | **ECG20** | **-335005.3** | **670050.6** | **670193.5** |
| Eukaryotes | WAG | -744073.4 | 1488149 | 1488157 |
| | LG | -737183.0 | 1474368 | 1474376 |
| | EX2 | -731952.0 | 1463908 | 1463925 |
| | EX3 | -730975.5 | 1461957 | 1461982 |
| | EHO | -733277.3 | 1466561 | 1466586 |
| | UL2 | -730422.0 | 1460848 | 1460865 |
| | UL3 | -727021.7 | 1454049 | 1454074 |
| | EX_EHO | -729388.5 | 1458789 | 1458839 |
| | ECG6 | -730336.5 | 1460685 | 1460735 |
| | ECG10 | -728197.6 | 1456415 | 1456498 |
| | **ECG20** | **-725660.6** | **1451361** | **1451528** |
| Metazoa | WAG | -852415.0 | 1704832 | 1704840 |
| | LG | -843611.9 | 1687226 | 1687234 |
| | EX2 | -834656.2 | 1669316 | 1669333 |
| | EX3 | -832621.5 | 1665249 | 1665274 |
| | EHO | -835282.8 | 1670572 | 1670597 |
| | UL2 | -831122.8 | 1662250 | 1662266 |
| | UL3 | -825262.4 | 1650531 | 1650556 |
| | EX_EHO | -830043.8 | 1660100 | 1660150 |
| | ECG6 | -830737.5 | 1661487 | 1661537 |
| | ECG10 | -825936.5 | 1651893 | 1651976 |
| | **ECG20** | **-821677.2** | **1643394** | **1643561** |

Table 2: **Goodness of fit on phylogenomics datasets.**

between them or to LG exchangeabilities than equilibrium frequencies are. Nonetheless, the weak differences of amino acid exchangeabilities among categories impact greatly on the likelihood Le et al. (2008b). It suggests that site specificities also concern exchangeabilities such

that amino acids that are more specific to a particular biological context (secondary structure, solvent exposure, etc) have both higher frequencies and exchangeabilities in the category describing this context. This suggests the development of mixture of matrices with more categories than the present mixtures (the EX_EHO model contains 6 categories) to further increase the fit to the data. But accounting for exchangeability heterogeneities leads to the estimation and use of many parameters (190 exchangeabilities per category) such that the estimation of high dimension mixtures of matrices (of the size of present mixtures of profiles) may lead to optimization issues and redundancy between the categories. Results presented in this study are in line with previous statements (Lartillot and Philippe, 2004; Lartillot et al., 2007; Le et al., 2008a) supporting the idea that having a reduced number of categories (of the size of the present mixtures of matrices) is not enough to faithfully model heterogeneities of sequence evolution. ECG models were proved to the be the best empirical mixture models currently available in the literature. As a sufficient number of profiles is mandatory for the model to accomodate multiple substitutions and so, to be more robust to phylogenetic artifacts (Lartillot et al., 2007), we think that efficient mixtures of profiles of the type of ECGs will have a strong impact on phylogenetic reconstructions and be relevant to phylogeneticists.

## References


Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* pages 267–281. Petrov BN, Csaki F, editors Budapest (Hungary).

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, and Sonnhammer ELL. 2002. The Pfam protein families database. *Nucleic Acids Res* 30:276–280.

Boussau B, Blanquart S, Necsulea A, Lartillot N, and Gouy M. 2008. Parallel Adaptation to High Temperature in the Archaean Eon. *Nature* 456:942–945.

Burnham KP and Anderson DR. 2004. Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods and Research* 33:261–304.

Douzery EJP, Snell EA, Bapteste E, Delsuc F, and Philippe H. 2004. The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils? *Proc Natl Acad Sci U S A* 101:15386–15391.

Drummond AJ, Ho SY, Phillips MJ, and Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *Plos Biol* 4:e88.

Dutheil J and Boussau B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol Biol* 8:255.

Dutheil JY, Galtier N, Romiguier J, Douzery EJP, Ranwez V, and Boussau B. 2012. Efficient Selection of Branch-Specific Models of Sequence Evolution. *Mol Biol Evol* 29:1861–1874.

Gascuel O and Guindon S. 2007. Modelling the variability of evolutionary processes. In *Reconstructing evolution: new mathematical and computational advances* pages 65–99. Oxford University Press Oxford, UK o. gascuel & m. steel edition.

Gaucher EA, Thomson JM, Burgan MF, and Benner SA. 2003. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature* 425:285–288.

Goldman N, Thorne JL, and Jones DT. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* 149:445–458.

Groussin M and Gouy M. 2011. Adaptation to Environmental Temperature Is a Major Determinant of Molecular Evolutionary Rates in Archaea. *Mol Biol Evol* 28:2661–2674.

Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, Bigot T, Fournier D, Pouyet F, Cahais V, Bernard A, Scornavacca C, Nabholz B, Haudry A, Dachary L, Galtier N, Belkhir K, and Dutheil JY. 2013. Bio++: Efficient Extensible Libraries and Tools for computational molecular evolution. *Mol Biol Evol* 30:1745–1750.

Harms MJ and Thornton JW. 2010. Analyzing protein structure and function using ancestral gene reconstruction. *Curr Opin Struct Biol* 20:360–366.

Holmes I and Rubin GM. 2002. An expectation maximization algorithm for training hidden substitution models. *J Mol Evol* 317:753–764.

Jones DT, Taylor WR, and Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8:275–282.

Koshi JM and Goldstein RA. 1998. Models of natural mutations including site heterogeneity. *Proteins* 32:289–295.

Lartillot N, Brinkmann H, and Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol* 7:S4.

Lartillot N, Lepage T, and Blanquart S. 2009. PhyloBayes 3. A Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25:2286–2288.

Lartillot N and Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol* 21:1095–2004.

Le SQ and Gascuel O. 2008. An Improved General Amino Acid Replacement Matrix. *Mol Biol Evol* 25:1307–1320.

Le SQ and Gascuel O. 2010. Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. *Syst Biol* 59:277–287.

Le SQ, Gascuel O, and Lartillot N. 2008a. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24:2317–2323.

Le SQ, Lartillot N, and Gascuel O. 2008b. Phylogenetic mixture models for proteins. *Phil. Trans. R. Soc. Lond. B* 363:3965–3976.

Liò P and Goldman N. 1999. Using protein structural information in evolutionary inference: transmembrane proteins. *Mol Biol Evol* 16(12):1696–1710.

Nielsen R and Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148(3):929–936.

Pagel M and Meade A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst Biol* 53:571–581.

Philippe H, Derelle R, Lopez P, Pick K, Borchiellini C, Bourry-Esnault N, Vacelet J, Renard E, Houliston E, Quéinnec E, Silva CD, Wincker P, Guyader HL, Leys S, Jackson DJ, Schreiber F, Erpenbeck D, Morgenstern B, Wörheide G, and Manuel M. 2009. Phylogenomics revives traditional views on deep animal relationships. *Curr Biol* 19:706–712.

Schneider R, de Daruvar A, and Sander C. 1997. The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res* 25:226–230.

Sugiura N. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics, Theory and Methods* A7:13–26.

Thorne JL, Goldman N, and Jones DT. 1996. Combining protein evolution and secondary structure. *Mol Biol Evol* 13(5):666–673.

Thorne JL, Kishino H, and Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol* 15:1647–1657.

Whelan S and Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* 18:691–699.

Yang Z. 2006. *Computational Molecular Evolution*. Oxford University Press oxford university press inc., new york edition.

Yang Z and Rannala B. 2006. Bayesian Estimation of Species Divergence Times Under a Molecular Clock Using Multiple Fossil Calibrations with Soft Bounds. *Mol Biol Evol* 23:212–226.

## 2.3 Modélisation conjointe de l'hétérogénéité entre sites et dans le temps.

### 2.3.1 Introduction

Comme expliqué dans les deux précédents articles, le processus évolutif varie le long de la séquence protéique et entre les lignées. Ces deux articles ont présenté deux types de modèles : COaLA qui est hétérogène en temps mais homogène en sites et ECG qui est homogène en temps mais hétérogène en sites. En 2008, Blanquart and Lartillot (2008) ont publié le premier modèle hétérogène en temps et en sites (CAT-BP), qui est capable de moduler les processus spécifiques des sites dans le temps, le long de l'arbre phylogénétique. Ce modèle a été implémenté dans le cadre bayesien et utilise les spécificités du modèle CAT (Lartillot and Philippe, 2004), hétérogène en sites et BP (Blanquart and Lartillot, 2006), hétérogène en temps. Ainsi, le modèle autorise l'ensemble des profils specifiques des sites de varier de manière corrélée selon l'endroit de l'arbre. CAT-BP optimise le nombre et la localisation des points de rupture le long de l'arbre où ces variations de composition globale se produisent, comme dans le modèle BP. Si l'on considère une zone donnée de l'arbre, les fréquences d'équilibres de chaque processus markoviens du modèle CAT, appelés par la suite profil, sont modulées par des modulateurs spécifiques de la zone considérée. Ainsi, le profil $\pi^{j,n}$ de la catégorie $j$ du modèle CAT dans la zone $n$ de l'arbre est calculé en multipliant le profil $\Pi^c_j$ spécifique de la catégorie $j$ par le modulateur $\Pi^m_n$ spécifique de la zone $n$, de telle sorte que, en utilisant le produit terme à terme des vecteurs :

$$\pi^{j,n} = \frac{1}{Z} \times \Pi^c_j \times \Pi^m_n$$

avec Z le facteur de normalisation égal à :

$$Z = \sum_{s=1}^{I} \Pi^c_{j,i} \times \Pi^m_{n,i}$$

et $I$ la taille de l'alphabet (20 dans le cas protéique).

Un tel modèle en Maximum de Vraisemblance n'est actuellement pas disponible dans la littérature. Ce que je présente dans cette partie est pour la première fois, **en Maximum de Vraisemblance**, un modèle hétérogène en temps et en sites qui permet de combiner l'approche COaLA avec tout type de modèle de mélange protéique. Ce modèle s'inspire grandement du modèle CAT-BP de telle sorte que les profils de chaque processus markovien du mélange sont modulés sur chaque branche par des modulateurs optimisés dans le même cadre mathématique que le modèle COaLA. Les premiers résultats de fit aux données sont encourageants et montrent

l'intérêt d'un tel modèle, nommé COaLAmix, afin de mieux capter les subtilités du processus évolutif agissant au niveau des séquences.

## 2.3.2 Matériels et Méthodes

### 2.3.2.1 Présentation du modèle

Le modèle COaLA considère que la variation de composition a lieu d'une branche à l'autre. Si l'on considère une branche donnée, cette branche est caractérisée par ses propres fréquences d'équilibres. COaLA calcule une analyse factorielle des correspondences (COA) à partir de la matrice contenant pour chaque espèce les fréquences observées en acides aminés de l'alignement.

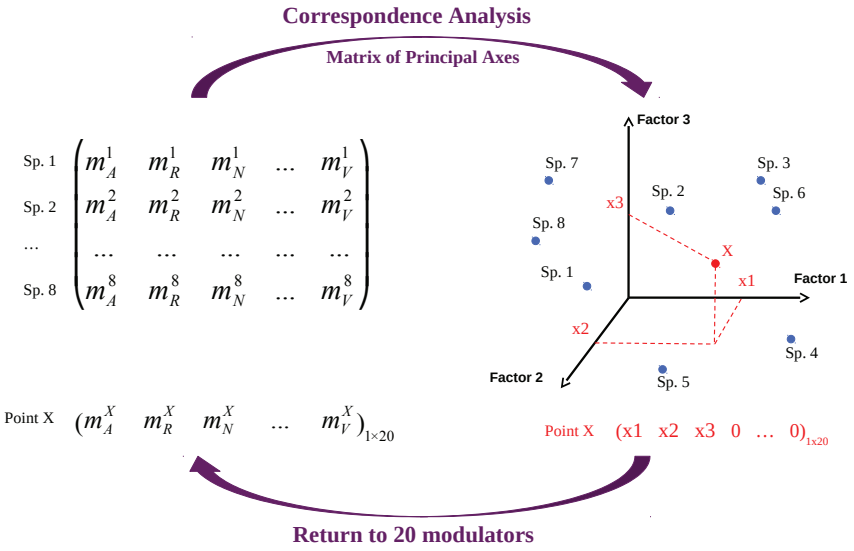



FIGURE 2.2 – Réduction de l'espace des paramètres à optimiser effectuée par le modèle COaLAmix.

COaLA explore alors pour chaque branche l'espace des fréquences d'équilibres au travers de l'espace de dimension moindre défini par les premiers axes de la COA. Au lieu de considérer la matrice des fréquences observées pour calculer directement la COA, COaLAmix utilise cette matrice pour en calculer une autre contenant les modulateurs observés. Un modulateur observé

$m_i^s$ pour un acide aminé $i$ donné chez une espèce $s$ est défini de la façon suivante :

$$m_i^s = \frac{f_i^s}{\frac{1}{S}\sum_s f_i^s}$$

avec S le nombre total d'espèces.

À partir de la matrice des modulateurs, une COA est calculée et est utilisée de la même façon que dans le modèle COaLA. À une branche donnée $b$ correspond un point $X$ dans l'espace défini par les premiers axes de la COA et les coordonnées de $X$ sont utilisées pour calculer le point correspondant dans l'espace à 20 dimensions des modulateurs en renversant la COA (Figure 2.2 et 2.3).

Par la suite, ce vecteur à 20 modulateurs spécifiques de la branche va servir à moduler les profils du modèle de mélange afin de les rendre spécifiques de la branche $b$, dans le même état d'esprit que dans le modèle CAT-BP (Figure 2.3). Ainsi, on a $\pi_i^{j,b}$, la fréquence d'équilibre de l'acide aminé $i$ dans le profil $C_j$ de la branche $b$ qui est égale, à une constante de normalisation près, à :

$$\pi_i^{j,b} \propto \pi_i^j \times m_i^b$$

avec $m_i^b$ le modulateur de $i$ spécifique de la branche $b$ et $\pi_i^j$ la fréquence d'équilibre de $i$ dans le profil $C_j$.

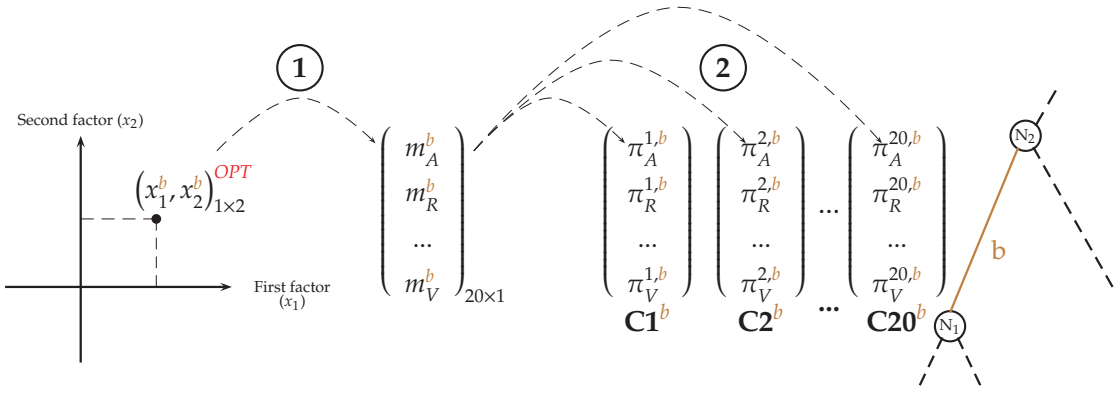



FIGURE 2.3 – Modulation des profiles selon les modulateurs spécifiques d'une branche de l'arbre phylogénétique.

Comme pour le modèle COaLA, COaLAmix optimise par Maximum de Vraisemblance les coordonnées du point $X$ dans l'espace réduit défini par la COA afin d'optimiser le vecteur des modulateurs de manière efficace.

#### 2.3.2.2 Jeu de données et expérience d'ajustement

Le jeu de données utilisé est le concaténat universel protéique construit par Boussau et al. (2008), contenant 30 espèces et 3336 sites et la topologie obtenue à partir de l'analyse de ce jeu de données et présentée en Figure 2 de Boussau et al. (2008). Le modèle hétérogène en temps et en sites a été comparé à d'autres modèles, soit homogènes en temps et en sites ou soit hétérogènes uniquement pour l'une des deux conditions. Les calculs ont été effectués avec bppML (Dutheil and Boussau, 2008) et le critère BIC a été utilisé pour déterminer le modèle qui s'ajuste le mieux aux données.

### 2.3.3 Résultats

Le tableau 2.1 montre les résultats d'ajustement aux données protéiques universelles de Boussau et al. (2008). Plusieurs modèles ont été comparés. Le modèle F81_C20 est simplement le modèle C20 proposé par Le et al. (2008a), qui utilise une matrice d'échangeabilités plate de type F81 et 20 profils, dont les fréquences d'équilibres sont empiriques. Au moment où j'ai testé la faisabilité du modèle COaLAmix, qui peut s'adapter à n'importe quel type de modèle de mélange, les modèles ECG présentés précédemment n'étaient pas encore disponibles. Comme expliqué en introduction de la section 2.2, les modèles C10 à C60 (Le et al., 2008a) ont un fit aux données relativement faible comparativement aux modèles de mélange de matrices de type UL3 (Le et al., 2008b). Nous avions donc décidé de tenter d'améliorer les modèles C10 à C60 existants en utilisant une matrice d'échangeabilités plus pertinante que la matrice F81. Dans un premier temps, nous avions ajouté aux modèles la matrice d'échangeabilités LG, commune à tous les profils du modèle C20. Les résultats étant toujours assez décevants vis à vis de UL3 sur le jeu de données de Boussau et al. (2008), nous avons tenté de ré-ajuster les valeurs d'échangeabilités de la matrice LG par rapport à chaque profil, afin d'obtenir un modèle de mélange de matrices. Le réajustement des échangeabilités proposé était le suivant :
En partant du principe que pour une catégorie donnée $c$, le taux instantané de la catégorie est égal au taux instantané de LG, on a

$$s_{ij}^{c}\pi_{j}^{c} = s_{ij}^{LG}\pi_{j}^{LG} \quad (1)$$

avec :

- $s_{ij}^{c}$ et $s_{ij}^{LG}$, l'échangeabilité de la catégorie $c$ et de LG, respectivement et

- $\pi_{j}^{c}$ et $\pi_{j}^{LG}$, les fréquences d'équilibres de $j$ de la catégorie $c$ et de LG, respectivement.

Ainsi,

$$s^c_{ij} = \frac{s^{LG}_{ij}\pi^{LG}_j}{\pi^c_j}. \quad (2)$$

Mais alors :

$s^c_{ij} \neq s^c_{ji}$, ce qui contredit la réversibilité du processus.

Finalement, $s^c_{ij}$ était calculé de cette façon :

$$s^c_{ij} = \frac{1}{2}\left[\frac{s^{LG}_{ij}\pi^{LG}_j}{\pi^c_j} + \frac{s^{LG}_{ij}\pi^{LG}_i}{\pi^c_i}\right] \text{ (3), de telle sorte que : } \quad s^c_{ij} = s^c_{ji}$$

Le tableau 2.1 montre que ce modèle de mélange de matrices basé sur les profils C20 et l'échangeabilité LG (LG_C20) s'ajuste très bien aux données d'après le critère BIC par rapport au modèle LG et par rapport au modèle F81_C20. Cependant, d'un point de vue conceptuel, ce ré-ajustement n'est pas optimal. Tout d'abord parce qu'il mélange des valeurs optimisées au maximum de vraisemblance (les profils C20) avec des valeurs re-calculées à partir d'une matrice d'échangeabilités (LG) dont les valeurs ont été optimisées indépendamment. Ensuite, la première hypothèse stipulant que $s^c_{ij}\pi^c_j = s^{LG}_{ij}\pi^{LG}_j$ va, pour un couple d'acide aminés donnés, rapprocher tous les taux de substitution instantanés des profiles vers le taux instantané de la matrice LG, et donc diminuer la specificité de chaque profil. En fait, l'hétérogénéité restante entre les profils vis à vis du taux de substitution instantané est uniquement dûe à la façon dont les nouvelles échangeabilités sont calculées (équation (3)) par la moyenne arithmétique. Cependant, cette approche permet de montrer qu'apporter de l'information sur les échangeabilités est primordial pour atteindre de bonnes performances en terme d'ajustement.

Les premiers tests réalisés avec le modèle COaLAmix ont été faits avec ce modèle LG_C20. Ils montrent (Tableau 2.1) que COaLAmix est capable d'ajuster extrêmement bien les données, et de surpasser les modèles hétérogènes en sites ou temps uniquement.

### 2.3.4 Discussion

Ces premiers résultats sont très encourageants. Cependant, quelques problèmes d'optimisation ont été remarqués avec l'utilisation du modèle COaLAmix. Il semble que le modèle soit plus sujet à des problèmes de maxima locaux et qu'il soit difficile pour les algorithmes d'optimisation d'effectuer une exploration efficace des paramètres. Les résultats présentés dans le tableau 2.1 ont été obtenus en utilisant un optimisation par l'algorithme de Brent implémenté dans Bio++ (Dutheil et al., 2006; Guéguen et al., 2013) sur chacun des paramètres. Mais si l'on utilise un algorithme de Newton-Raphson, les vraisemblances finales sont systématiquement moins bonnes,

| | Universal data set | | |
|---|---|---|---|
| **Processus** | **Modèle** | **ΔlnL** | **ΔBIC** |
| Homogeneous | LG | 0 | 0 |
| Hétérogène en sites | F81_C20 | +1458 | -2761 |
| | LG_C20 | +2328 | -4502 |
| Hétérogène en temps | LG+COaLA[1] | +435 | -392 |
| | LG+COaLA[2] | +688 | -419 |
| | LG+COaLA[3] | +977 | -518 |
| Hétérogène en temps et sites | LG_C20+COaLAmix[1] | +2849 | -5066 |
| | LG_C20+COaLAmix[2] | +3351 | -5591 |
| | **LG_C20+COaLAmix[3]** | **+3687** | **-5784** |

TABLEAU 2.1 – Efficacité du modèle COaLAmix vis à vis du fit aux données.

montrant que l'optimisation reste coincée dans l'espace des paramètres. Des tests doivent encore être effectués avec des méthodes de gradient pour estimer la difficulté d'exploration de l'espace des paramètres et plus globalement, pour essayer de rendre le modèle plus exploitable en termes d'efficacité d'optimisation.

Les résultats du tableau 2.1 montrent l'importance de ne pas négliger ces deux types d'hétérogénéité du processus évolutif, tous les deux à l'oeuvre lors de l'évolution des séquences protéiques. Il est aussi intéressant de remarquer qu'il semble y avoir un effet synergique de la co-modélisation de ces deux processus à la fois, puisque le gain de vraisemblance ou de fit obtenu avec le modèle COaLAmix n'est pas la simple addition des deux gains apportés par chacune des deux composantes du modèle. Il sera intéressant de voir si cet effet se retrouve sur d'autres jeux de données.

# 3

# Renaissance *in silico* et évolution précoce du monde microbien.

## 3.1 Température ancestrale de vie des Archées et leurs taux d'évolution.

### 3.1.1 Introduction

Chez les Bactéries et les Archées, de nombreux facteurs biologiques influencent la composition en bases des ARN structuraux (ARN ribosomique ou de transfert par exemple) et en acides aminés des protéines. Ainsi, la composition en G+C du génome est le facteur majeur expliquant la plus grande part de variabilité en acides aminés entre espèces. Le deuxième facteur majeur est la température, qui contraint les molécules d'ARN structuraux et les protéines à s'adapter afin de répondre à la contrainte physique exercée par la température. Cette adaptation à basse ou haute température entraine une variation de composition moléculaire, afin d'optimiser la stabilité et la fonction des molécules selon les conditions de vie de l'organisme en terme de température. Cette relation forte entre composition et température de vie a permis à Galtier et al. (1999) d'estimer quelle étaient la température de vie du dernier ancêtre commun universel (LUCA) à partir

de l'estimation des compositions moléculaires ancestrales. Par la suite, Boussau et al. (2008) a également utilisé cette approche afin d'étudier l'évolution de l'adaptation à la température le long de l'arbre de la vie. Cette section s'intéresse tout particulièrement à l'adaptation à la température de vie le long de l'arbre des Archées. Comme démontré dans le chapitre précédent, les modèles non-homogènes dans le temps, autorisant une variation compositionnelle entre lignées, sont nécessaires afin d'estimer précisemment les compositions ancestrales. De tels modèles ont été utilisés dans l'article présenté ici afin d'estimer les compositions moléculaires ancestrales des Archées et de les relier à leur température ancestrales. Cette étude a notamment permis de montrer que l'adaptation à la température est une des raisons majeures expliquant la violation de l'hypothèse d'horloge moléculaire stipulant une constance des taux entre lignées.

Cet article a été publié dans le journal Molecular Biology & Evolution.

### 3.1.2 Manuscrit

Molecular Biology and Evolution

Research article

# Adaptation to Environmental Temperature Is a Major Determinant of Molecular Evolutionary Rates in Archaea

Mathieu Groussin*[,1] and Manolo Gouy[1]

[1]Laboratoire de Biométrie et Biologie Evolutive, CNRS, Université de Lyon, Université Lyon I, Villeurbanne, France

***Corresponding author:** E-mail: mathieu.groussin@etu.univ-lyon1.fr.

**Associate editor:** Hervé Philippe

## Abstract

Methods to infer the ancestral conditions of life are commonly based on geological and paleontological analyses. Recently, several studies used genome sequences to gain information about past ecological conditions taking advantage of the property that the G+C and amino acid contents of bacterial and archaeal ribosomal DNA genes and proteins, respectively, are strongly influenced by the environmental temperature. The adaptation to optimal growth temperature (OGT) since the Last Universal Common Ancestor (LUCA) over the universal tree of life was examined, and it was concluded that LUCA was likely to have been a mesophilic organism and that a parallel adaptation to high temperature occurred independently along the two lineages leading to the ancestors of Bacteria on one side and of Archaea and Eukarya on the other side. Here, we focus on Archaea to gain a precise view of the adaptation to OGT over time in this domain. It has been often proposed on the basis of indirect evidence that the last archaeal common ancestor was a hyperthermophilic organism. Moreover, many results showed the influence of environmental temperature on the evolutionary dynamics of archaeal genomes: Thermophilic organisms generally display lower evolutionary rates than mesophiles. However, to our knowledge, no study tried to explain the differences of evolutionary rates for the entire archaeal domain and to investigate the evolution of substitution rates over time. A comprehensive archaeal phylogeny and a non homogeneous model of the molecular evolutionary process allowed us to estimate ancestral base and amino acid compositions and OGTs at each internal node of the archaeal phylogenetic tree. The last archaeal common ancestor is predicted to have been hyperthermophilic and adaptations to cooler environments can be observed for extant mesophilic species. Furthermore, mesophilic species present both long branches and high variation of nucleotide and amino acid compositions since the last archaeal common ancestor. The increase of substitution rates observed in mesophilic lineages along all their branches can be interpreted as an ongoing adaptation to colder temperatures and to new metabolisms. We conclude that environmental temperature is a major factor that governs evolutionary rates in Archaea.

**Key words:** Archaea, evolutionary rates, optimal growth temperature, ancestral sequence reconstruction, nonhomogeneous models.

## Introduction

Bacteria and Archaea show adaptations to many kinds of environments and especially to a wide range of temperatures. Several recent studies have attempted to reconstruct ancestral environmental temperatures using molecular sequence data (Galtier et al. 1999; Boussau and Gouy 2006; Boussau et al. 2008; Gaucher et al. 2008). These analyses exploited the signal left by environmental temperature on both extant and ancestral sequences in terms of base and amino acid compositions. Boussau et al. (2008) studied the evolution of thermophily along the universal tree of life using two molecular thermometers based on the compositions of ribosomal RNA (rRNA) and protein sequences. They concluded firstly that LUCA (the Last Universal Common Ancestor) lived at low temperatures, secondly that parallel adaptations to high temperatures occurred from LUCA to the last common ancestor of Bacteria and to that of Archaea, and thirdly that optimal growth temperatures (OGTs) decreased with time in the bacterial domain. These results were obtained with 30 organisms, among which only seven archaeal species. This limitation prevented a precise study of the evolution of OGT in the archaeal domain. Presently available fully sequenced archaeal genomes give the opportunity to investigate in greater detail the evolutionary history of OGT in this domain.

It has long been observed that thermophilic lineages tend to have shorter branches in archaeal phylogenetic trees than do mesophilic lineages (Stetter 2006). Several factors have been proposed to explain why molecular evolutionary rates vary between organisms (Bromham 2009). In vertebrates, the generation time is critical in the determination of evolutionary rates (Bromham et al. 1996). In mammals, it has been shown that population size, body size, and metabolic rates are probably involved in shaping molecular evolutionary rates (Bromham 2009). Concerning Archaea and Bacteria, few factors are known to explain the differences of evolutionary rates between species. Nevertheless, it seems clear that for all species, and particularly in Bacteria where it has been shown, the efficiency of DNA replication and DNA repair machineries is under selection and determines substitution rates (Denamur and Matic 2006). Archaeal and bacterial species are considerably sensitive to the variations of their environment and to

the variations of mutagen concentrations (e.g., UV, temperature) (Foster 2007). Thus, Valentine (2007) proposed that chronic energy stress, from a metabolic and thermodynamic point of view, is the major selective pressure that governs evolutionary rates in Archaea. A physical factor that could cause such energy stress is environmental temperature. Thus, previous studies showed that thermophilic species are characterized by a stronger purifying selection than mesophiles. Indeed, Friedman et al. (2004) showed that thermophiles display a lower ratio of nonsynonymous to synonymous substitutions than mesophiles. This is consistent with the idea that proteins of species living in hot environments are more functionally constrained (Vetriani et al. 1998). Drake suggested that thermophiles exhibit very low genomic mutation rates and that this phenomenon could be explained by an adaptation to avoid deleterious mutations at high temperatures (Drake 2009). However, most studies that focused on mutational and evolutionary rates in Archaea or in thermophiles were restricted to few species (Grogan et al. 2001; Friedman et al. 2004; Mackwan et al. 2007, 2008; Drake 2009) making it impossible to have a vision at the scale of the entire domain. In this work, we attempt to reconstruct the evolutionary history of environmental temperatures at the level of the entire archaeal domain and investigate whether there is evidence that evolutionary rates are constrained by environmental temperatures.

The reconstruction of ancestral environmental temperatures and evolutionary rates using extant molecular data requires statistical models of the molecular evolutionary process and a phylogenetic tree of the organisms under study. Phylogenetic relationships between all archaeal species remain debated. Two major phyla have long been recognized (Gribaldo and Brochier-Armanet 2006): Euryarchaea, which is composed of thermoacidophiles, methanogens, extreme halophiles, and a few hyperthermophiles, and Crenarchaea, which were believed to be restricted to hyperthermophiles until mesophilic crenarchaeal species were discovered (DeLong 1992). These mesophilic species were grouped with Crenarchaea on the basis of 16S rRNA phylogenies. Brochier-Armanet et al. (2008) recently questioned the dichotomy between Euryarchaea and Crenarchaea in an analysis of *Cenarchaeum symbiosum*, the first mesophilic crenarchaeon entirely sequenced. They proposed that this group of mesophilic organisms should not be considered as Crenarchaea but rather as a third phylum, named Thaumarchaea, which diverged first in the archaeal tree. However, this conclusion remains uncertain. The evolutionary origins of other archaeal species are also unresolved, for example, that of the recently sequenced *Candidatus Korarchaeum cryptofilum* (Elkins et al. 2008).

We used here nonhomogeneous evolutionary models, which have been shown to be more realistic than homogeneous models (Dutheil and Boussau 2008). These models were used to infer base and amino acid compositions at each ancestral node of the archaeal tree. Through the use of appropriate molecular thermometers, these compositions permitted us to deduce OGT along the tree. We inferred that the ancestors of Archaea, Crenarchaea, and Euryarchaea were hyperthermophiles, and therefore, that ancestral archaeal species were adapted to hot environments. Furthermore, a strong relationship between environmental temperature and molecular evolutionary rates in Archaea has been identified. This implies that environmental temperature has been a major determinant of evolutionary rates in the archaeal domain.

## Materials and Methods

### Data Retrieval

Thirty-five completely sequenced archaeal genomes were selected for the phylogenetic studies to represent all known archaeal biodiversity. However, all 56 genomes available in GenBank as of February 2009 were used to construct the rRNA and protein data sets (see rRNA and Protein Data Sets). Protein sequences were downloaded from the Hogenom database (Penel et al. 2009) when possible. The genomes of *Ignicoccus hospitalis*, *Desulfurococcus kamchatkensis*, *Metallosphaera sedula*, *Caldivirga maquilingensis*, *Pyrobaculum arsenaticum*, *Thermoproteus neutrophilus*, *Halobacterium salinarum*, *Methanococcus voltae*, *Methanobrevibacter smithii*, *C. Korarchaeum cryptofilum* and *Nitrosopumilus maritimus* were retrieved from GenBank database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Two thaumarchaeal fosmids sequences, also extracted from GenBank, were added to the study to increase the diversity within this group: *uncultured crenarchaeote 74A4* and *uncultured crenarchaeote KM3-34-D9*. For these two mesophilic species, both small and large rRNA subunits were used in the rRNA data set, whereas only the elongation factor G protein for *KM3-34-D9* and the 30S ribosomal protein S10 for *74A4* were available and used in the protein data set.

All bacterial and eukaryal genomes were retrieved from the Hogenom database, with the exception of the *Giardia lamblia* genome, which was extracted from the GiardiaDB database (http://giardiadb.org/giardiadb/). The complete list of genomes with their origin is in supplementary table 1 (Supplementary Material online).

### rRNA and Protein Data Sets

The rRNA and protein alignments were constructed as follows: the 56 species were used in order to improve the quality of the alignment with as much information as possible. Then, species not exploited in the following steps were removed from the final alignments, which contain 35 species. The small and the large subunits (SSUs and LSUs) of archaeal, bacterial, and eukaryal rRNAs were extracted from GenBank. Archaeal rRNA SSUs and LSUs were aligned separately with the Silva aligner (http://www.arb-silva.de/aligner/), which takes secondary structures into account. Bacterial rRNAs were aligned following the same procedure. Then, archaeal rRNAs were aligned with bacterial rRNAs, using the "profile alignment" function of the ClustalW program (Thompson et al. 2002). For the data sets containing the three domains

of life, archaeal and bacterial rRNAs were first aligned together and then aligned by profile against eukaryal rRNAs with ClustalW. Eukaryal 5.8S rRNAs were added upstream from large eukaryal subunits because they are homologous to the 5′ end of the large prokaryotic subunits (Nazar 1980). Then, SSUs and LSUs rRNAs were concatenated with ScaFos (Roure et al. 2007). Fast-evolving sites were subsequently removed from the alignment by the Gblocks program, with standard options, allowing gap positions (Castresana 2000). The final archaeal + bacterial rRNA data set contains 3,719 sites. With the three domains, Gblocks retained 3,629 sites. Protein gene families extracted from the Hogenom database (Penel et al. 2009) were selected with different selection criteria. First, universal and single-copy gene families for all 56 archaeal genomes were retrieved. Second, gene families that are universal and single-copy only for Euryarchaea or Crenarchaea were also selected. Gene families affected by "distant" horizontal gene transfers (HGTs) were removed from the selection, distant HGTs being defined by topologies of single-gene phylogenies that do not respect the monophyly of Crenarchaea and Euryarchaea, as presented in the consensus archaeal phylogeny of Brochier-Armanet et al. (2008). *Nanoarchaeum equitans*, *C. Korarchaeum*, and the two thaumarchaeal species were not taken into account in this approach as their position is highly controversial (Brochier-Armanet et al. 2008). As a result, 72 gene families were conserved for the archaeal + bacterial phylogenies and 68 gene families for the universal tree (supplementary table 2, Supplementary Material online). We stress here that the aim of this study is to investigate the evolution of the adaptation to OGT in Archaea at the compositional level and not to completely solve the archaeal phylogeny. Indeed, it is likely that many gene families retained for this analysis are affected by HGT, but we hypothesize that these HGT did not shape the long-term evolution of proteins at the compositional level. Each family was aligned by Muscle (Edgar 2004) and treated by Gblocks (Castresana 2000) with standard parameters and all gaps allowed. Overall, 9,799 and 8,598 sites were retained for the two-domain and three-domain alignments, respectively.

## Phylogenetic Reconstructions

In Archaea, three taxa were defined a priori: Crenarchaea, Euryarchaea, and Thaumarchaea. The monophyly of these three phyla was strongly supported by the analysis of Brochier-Armanet et al. (2008). The TreeFinder program was used to resolve multifurcations within each of these taxa. As it is extremely difficult to place *N. equitans* (its genome is highly degenerated because of its parasitic way of life (Hubert et al. 2002; Forterre et al. 2009)), it was deliberately placed within Euryarchaea, based on previous results (Brochier et al. 2006). For rRNAs, the GTR+$\Gamma_8$+I model was used. Concerning proteins, the LG substitution model was employed (Le and Gascuel 2008) with a gamma law (four categories). No proportion of invariant was considered (this proportion was at first estimated and was revealed to be negligible). Bootstrap analysis was computed with PhyML (Guindon and Gascuel 2003) (100 replicates). To reduce risks of long-branch attraction, PhyML-CAT (Le et al. 2008) was used to confirm the protein results obtained with PhyML. We chose 20 profiles (model C20) and applied a gamma law (four categories).

## Nonhomogeneous Models of Evolution

All nonhomogeneous experiments were carried out with BppML, belonging to the BppSuite of Programs (Dutheil and Boussau 2008). The following options were used: all sites were taken into account with no restriction on the percentage of gaps (maximum amount of allowed gaps of 100%) and all root frequencies were initially set to one per size of the alphabet (4 for RNA, 20 for proteins). A gamma law was added to all models that were tested, with eight and four categories for rRNAs and proteins, respectively. A proportion of invariants was also considered for rRNAs. We chose a simple likelihood recursion with a recursive site compression. All other options were set to default values. BppML allowed to estimate evolutionary parameters such as substitution and rate distribution parameters, ancestral frequencies, and branch lengths from the reference topology, which remains fixed. Different models of substitutions have been tested: T92, HKY85, and GTR models for the rRNA data set and JTT92, WAG, and LG models for the protein data set. The aim of this process was to fit as well as possible the compositional heterogeneity of the data set and to improve the estimation of evolutionary parameters (e.g., branch lengths). For the nonhomogeneous approach, we defined for each model several submodels in which parameters are either shared by the whole tree or assigned to one branch or to a specific group of branches. Three approaches have been used. The first assigned one substitution model per branch. The second approach assigned one substitution model to each phylum (Crenarchaea, Euryarchaea, Thaumarchaea, and Korarchaea), plus one to Bacteria (and one to Eukarya when present). The third approach considered again each phylum separately. However, inside each phylum, a further distinction between thermophiles and mesophiles has been added to the model. Concerning the universal tree, one specific model has been assigned to the GC-rich *G. lamblia* species.

The inference of ancestral rRNA and protein sequences at each node of the tree was performed by bppAncestor with previously computed parameters (Dutheil and Boussau 2008). Concerning rRNAs, BppML was first run with the whole alignment (3,719 sites or 3,629 for the two-domain or the three-domain alignment, respectively); the reconstruction of ancestral sequences was performed with an rRNA data set restricted to double-stranded regions. This second rRNA data set (1,801 sites or 1,142 sites for the two-domain or the three-domain alignment, respectively) was obtained by eliminating single-stranded sites manually with SeaView (Gouy et al. 2010). One hundred ancestral sequences for each node of the tree were inferred, and their average G+C content or amino acid composition were computed. Confidence

intervals (95%) of ancestral OGTs were computed following Boussau et al. (2008) with 100 bootstrap replicates.

### Statistics

Statistical computations were performed using R (http://www.R-project.org). Multivariate analyses were realized using the ade4 package (Thioulouse et al. 1997). All correlation coefficients presented in this study are statistically different from zero. The phylogenetic independent contrasts (PICs) analysis was performed using the ape package (Paradis et al. 2004). Bowker's tests were computed with the R scripts made available by Ababneh et al. (2006) at http://www.maths.usyd.edu.au/u/johnr/testsym/.

### Optimal Growth Temperatures

OGTs of Bacteria and Archaea were extracted from the German National Resource Centre for Biological Material (DSMZ, http://www.dsmz.de/). We referred to the literature for two bacteria, *Pseudomonas entomophila* and *Anaeromyxobacter dehalogenans*, because the DSMZ database does not provide such data (He and Sanford 2003; Hegan et al. 2007). Following Boussau et al. (2008), we defined three temperature classes: mesophiles when OGT $\leq$ 50 °C, thermophiles when OGT is between 50 °C and 80 °C, and hyperthermophiles when OGT $\geq$ 80 °C. The complete list of OGTs for all species is available in supplementary table 1 (Supplementary Material online).

## Results

### Domain-Scale Archaeal Phylogeny

Before inferring ancestral archaeal sequences and compositions to study the evolution of thermophily within Archaea, a phylogenetic reconstruction was carried out to obtain a reliable topology. Previous results (Brochier-Armanet et al. 2008; Elkins et al. 2008) revealed some uncertainty concerning the positions in the archaeal tree of key species in respect to the adaptation to OGT, like Thaumarchaea (originally named mesophilic crenarchaea) and the recently sequenced *C. Korarchaeum cryptophilum*. Cox et al. (2008) and Foster et al. (2009) recently questioned the monophyly of Archaea with phylogenetic results supporting the grouping of Eukarya and Crenarchaea to the exclusion of Euryarchaea. Lake et al. (1984) first proposed this hypothesis, known as the "eocyte" hypothesis. Therefore, Bacteria were chosen as the outgroup of Archaea to build rRNA and protein trees. Assuming the monophyly of Crenarchaea and Euryarchaea, we defined four major archaeal groups (Euryarchaea, Crenarchaea, Thaumarchaea, and Korarchaea)—because major topology ambiguities concern the positions of these phyla—and explored the 15 topologies defined by all possible arrangements between these groups (supplementary fig. 1, Supplementary Material online). We identified the best topology supported by RNA and protein data, based on the sum of rRNA and protein log-likelihoods and results of the expected-likelihood weights (ELW) statistical test (Strimmer and Rambaut 2002). Likelihoods for each possible topology were estimated with TreeFinder (Jobb et al. 2004), which optimized the tree within each predefined group.

Supplementary figure 1 (Supplementary Material online) summarizes the results and reveals that topology no. 14, where Thaumarchaea are a sister group of Korarchaea, both of them being a sister group of Crenarchaea, possesses the best maximum likelihood for both rRNAs and proteins. The ELW test cannot statistically rule out topologies no. 13 and 15. In these two topologies, the affinity between Thaumarchaea and Korarchaea disappears but Thaumarchaea never branch deeply in the tree. This maximum likelihood topology was also found using PhyML-CAT (Le et al. 2008) (which implements a rough approximation of the site-heterogeneous mixture model CAT (Lartillot and Philippe 2004)). The CAT model has been proven to be less sensitive to long-branch attraction (Lartillot et al. 2007) and confirmed the affinity between Thaumarchaea and Korarchaea. Several positions of Thaumarchaea have been proposed so far, for instance, a deep branching in the archaeal tree (Brochier-Armanet et al. 2008) or a branching within Crenarchaea (Elkins et al. 2008). Our results do not support these hypotheses but do not allow ruling them out because our protein data set is likely to be affected by HGT (see Material and Methods and Discussion). However, rRNA and protein phylogenies converge toward the same tree of the four predefined groups, and topologies within Crenarchaea and Euryarchaea are very similar for the rRNA and protein data sets. In order to check whether our results are robust to uncertainties in the phylogenetic tree of the archaeal domain, several alternative topologies were also used to infer ancestral OGTs (see The Influence of the Input Topology). To control whether the presence of eukaryotic sequences changes the inferences of ancestral OGTs in the archaeal domain (see below), the same approach was used to determine the best universal tree of life. Supplementary figure 1 (Supplementary Material online) shows that the eocyte topology (topology A), which clusters Eukarya and the association between Crenarchaea, Thaumarchaea, and Korarchaea, obtained the best maximum likelihood scores. This result is in agreement with recent propositions (Cox et al. 2008; Foster et al. 2009).

### Nonhomogeneous Modeling of the Molecular Evolution of Archaea

The chosen topology (no. 14 in supplementary fig. 1, Supplementary Material online) was used as input tree to run nonhomogeneous models of evolution implemented in the BppML program (Dutheil and Boussau 2008). See Material and Methods for a full description of the homogeneous and nonhomogeneous models used. Table 1 sums up all the results obtained with the HKY85 and LG models (The results obtained with the T92, GTR, JTT92, and WAG models are shown in supplementary table 3, Supplementary Material online.). Usually, models that are more parameter rich will have a higher likelihood than a more restricted model (Felsenstein 2004). Here, as we have different models,

**Table 1.** Estimation of the Best Nonhomogeneous Model of Evolution for rRNAs and Proteins.

| Substitution Model | Model Attribution | Parameters Shared in the Whole Tree | Number of Parameters | ln L | BIC values |
|---|---|---|---|---|---|
| HKY85 (rRNAs) | Homogeneous | All | 7 | −69578 | 139188.9 |
| | Per Branch | None | 355 | −67866 | 138650.5 |
| | | $\theta^a/\kappa^b$ | 105 | −69361 | 139585.2 |
| | | $\kappa$ | 268 | −68089 | 138381.3 |
| | | $\theta$ | 268 | −68955.7 | 140114.7 |
| | | $\theta_1{}^c/\theta_2{}^d$ | 181 | −67981.7 | 137451.4 |
| | | $\kappa/\theta_1/\theta_2$ | 94 | −68196.9 | 137166.6 |
| | | $\kappa/\theta_1$ | 181 | −68147.1 | 137782.2 |
| | | $\kappa/\theta_2$ | 181 | −68138.1 | 137764.2 |
| | Per Phylum | None | 27 | −69077.1 | 138376.2 |
| | | $\kappa$ | 22 | −69131 | 138442.9 |
| | | $\kappa/\theta_1/\theta_2$ | 12 | −69140.6 | 138379.9 |
| | | $\kappa/\theta_1$ | 17 | −69133.6 | 138407 |
| | | $\kappa/\theta_2$ | 17 | −69138 | 138415.8 |
| | Per group of extant species sharing similar OGT[e] | None | 59 | −68484.6 | 137454.3 |
| | | $\kappa$ | 46 | −68559.5 | 137497.2 |
| | | $\kappa/\theta_1/\theta_2$ | 20 | −68596.9 | 136358.2 |
| | | $\kappa/\theta_1$ | 33 | −68578.9 | 137429.1 |
| | | $\kappa/\theta_2$ | 33 | −68575.1 | 137421.5 |
| LG + F (Proteins) | Homogeneous | — | 20 | −447106 | 894395.8 |
| | Per Phylum | — | 134 | −446059 | 893349.5 |
| | Per group of extant species sharing similar OGT[e] | — | 324 | −444820 | 892617.7 |

[a] Equilibrium G+C content ($\pi G+\pi C$).
[b] Transition/transversion ratio.
[c] $\theta_1 = \pi A/(\pi A+\pi T)$.
[d] $\theta_2 = \pi G/(\pi G+\pi C)$.
[e] Within a phylum, a further distinction is made between mesophilic and thermophilic species for the attribution of the sets of equilibrium frequencies. The best model for rRNAs and proteins is in bold characters.

selecting the model that has the highest likelihood could lead to the choice of an unreasonably complex model. Thus, bayesian information criterion (BIC) tests have been carried out for each model to balance the effect of the number of parameters on the final likelihoods. BIC has been preferred to akaike information criterion because it penalizes more parameter-rich models (Ripplinger and Sullivan 2008), as occurs in this study. For rRNAs, the best BIC score is attained by the HKY85 model with $\kappa$ (transition/transversion ratio), $\theta_1$ ($= \pi\tilde{A}\pi A\pi T$), and $\theta_2$ ($= \pi G/(\pi G + \pi C)$) shared in the whole tree and one $\theta$ ($= \pi G + \pi C$) per branch. In the rest of the study, this model will be referred to as HKY85θpb (HKY85 with one $\theta$ per branch). It suggests that the HKY85θpb model represents a good compromise between the number of parameters and the ability to fit the data. Thus, homogeneous models do not fit properly the data because of their simplicity and, conversely, the assumption of one GTR model per branch is too complex and overparameterized (supplementary table 3, Supplementary Material online). As already mentioned by Dutheil and Boussau (2008), we did not observe a significant improvement of the results by allowing different $\kappa$ on each branch or group. Concerning the protein data set, the same approach has been performed, and we observed that the model with one set of LG-based equilibrium frequencies per group of mesophilic or thermophilic organisms obtained the best statistical scores among all tested models. Finally, we ran nonhomogeneous experiments for the 14 other topologies used as input trees with the two selected protein and rRNA models described above. Topology no. 14 remains the maximum likelihood topology with this nonhomogeneous approach (data not shown).

To assess if the HKY85θpb model properly fits the heterogeneity of the rRNA data set, we used a parametric bootstrapping method based on Bowker's test designed by Dutheil and Boussau (2008). This is a pairwise test that allows to detect whether two sequences evolved under two different processes (Ababneh et al. 2006). Bowker's tests have been performed for all rRNA sequence pairs. The number of significant Bowker's tests defined the heterogeneity of the alignment. We used the parameters of the HKY85θpb model that had been previously estimated by BppML to simulate 10,000 data sets with BppSeqGen (Dutheil and Boussau 2008). For each simulated alignment, the heterogeneity was calculated and the distribution of heterogeneity values for the whole simulated sequences was obtained. Finally, the heterogeneity value of the initial data set was compared with this distribution. Clearly, the

HKY85 homogeneous model does not fit the data because the simulated distribution significantly underestimates the heterogeneity of sequences (supplementary fig. 2, Supplementary Material online). However, the HKY85θpb model produced simulated sequences with heterogeneity values that are representative of the intrinsic heterogeneity of the original rRNA data set ($P$ value = 0.188) (supplementary fig. 2, Supplementary Material online).

## Inference of Ancestral OGTs

Previous studies proved that rRNA G+C content and OGT were strongly correlated in Bacteria and Archaea (Galtier and Lobry 1997) and used this correlation as a molecular thermometer to infer ancestral OGTs. To establish this relationship with our rRNA data set, we retained only double-stranded regions because it has been shown that equilibrium frequency estimations were biased toward frequencies at slowly evolving sites (single-stranded regions in rRNAs) when heterogeneous models of evolution are used (Gowri-Shankar and Rattray 2006). We obtained a double-stranded regions data set of 1,801 sites. The G+C content of these regions is highly correlated to OGT (fig. 1*A*, r = 0.95, $P$ value < 0.001). Concerning our protein data set, a correspondence analysis has been performed on the amino acid compositions of our alignment. This procedure, introduced by Boussau et al. (2008), produced another molecular thermometer. The results (fig. 1*B*) show that OGT and amino acid compositions are strongly linked together. Two major independent factors explain most of the variance in amino acid compositions in archaeal proteins. The first factor (41.5% of the total variance) highly correlates with genomic G+C content ($r$ = 0.81, $P$ value < 0.001) and the second factor (26.7% of the total variance) with OGT ($r$ = 0.84, $P$ value < 0.001).

However, the regressions of figure 1*A* and *B* are made with data points that are not statistically independent. Indeed, each data point being one species, the nonindependence arises from the fact that all species share a common ancestry and are not independently drawn from the same distribution. Thus, if a strong phylogenetic inertia exists in the traits under study, closely related species will tend to have similar values for the two traits and, consequently, will tend to cluster together in a regression diagram, increasing the correlation coefficient (Felsenstein 1985; Harvey and Pagel 1991). This problem has been noticed before, and several methods have been proposed to take the nonindependence of taxa into account (Lanfear et al. 2010). One of these methods, the PICs, proposed early on by Felsenstein (1985), has been employed here. The PIC approach uses the original values of each trait for all species and transforms them to produce new values, called contrasts, that are statistically independent and identically distributed and that can be compared by a correlation test. New correlation coefficients were calculated from the contrasts in OGT and in G+C content on one side ($r$ = 0.85, $P$ value < 0.001) and in second factor values ($r$ = 0.7, $P$ value < 0.001) on the other,

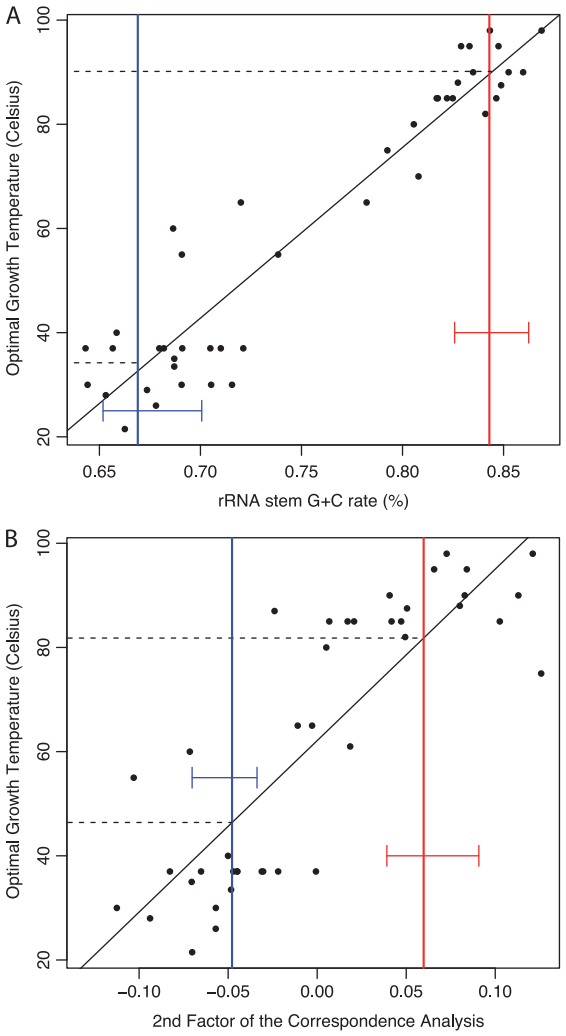



**Fig. 1.** Correlations between nucleotide or amino acid compositions and OGT. (*A*) rRNA thermometer. (*B*) Protein thermometer. In each plot, black dots indicate the positions of extant archaea and bacteria. For rRNAs, the linear correlation coefficient between OGT and rRNA stem G+C content is 0.95 ($P$ value < 0.001). For proteins, the second factor values of the correspondence analysis are strongly correlated with OGT ($r$ = 0.84, $P$ value < 0.001). Vertical lines represent the inferred compositions for the ancestor of Thaumarchaea (blue) and for the HACA (red) with their 95% confidence interval. Dashed lines represent the projection of ancestral compositions on the OGT axis. The HACA is predicted to be hyperthermophile, by both rRNAs and proteins.

confirming that the strong relationship between OGT and molecular compositions in rRNAs and proteins initially observed was not solely due to the nonindependence of data points.

Evolutionary model parameters initially estimated by BppML (e.g., branch lengths, substitution model parameters, gamma law parameter) were used by BppAncestor (Dutheil and Boussau 2008) to reconstruct rRNA and protein ancestral sequences, using the same topology. One hundred putative ancestral sequences were estimated for both data sets at each node of the tree. The G+C contents and amino acid compositions of these ancestral

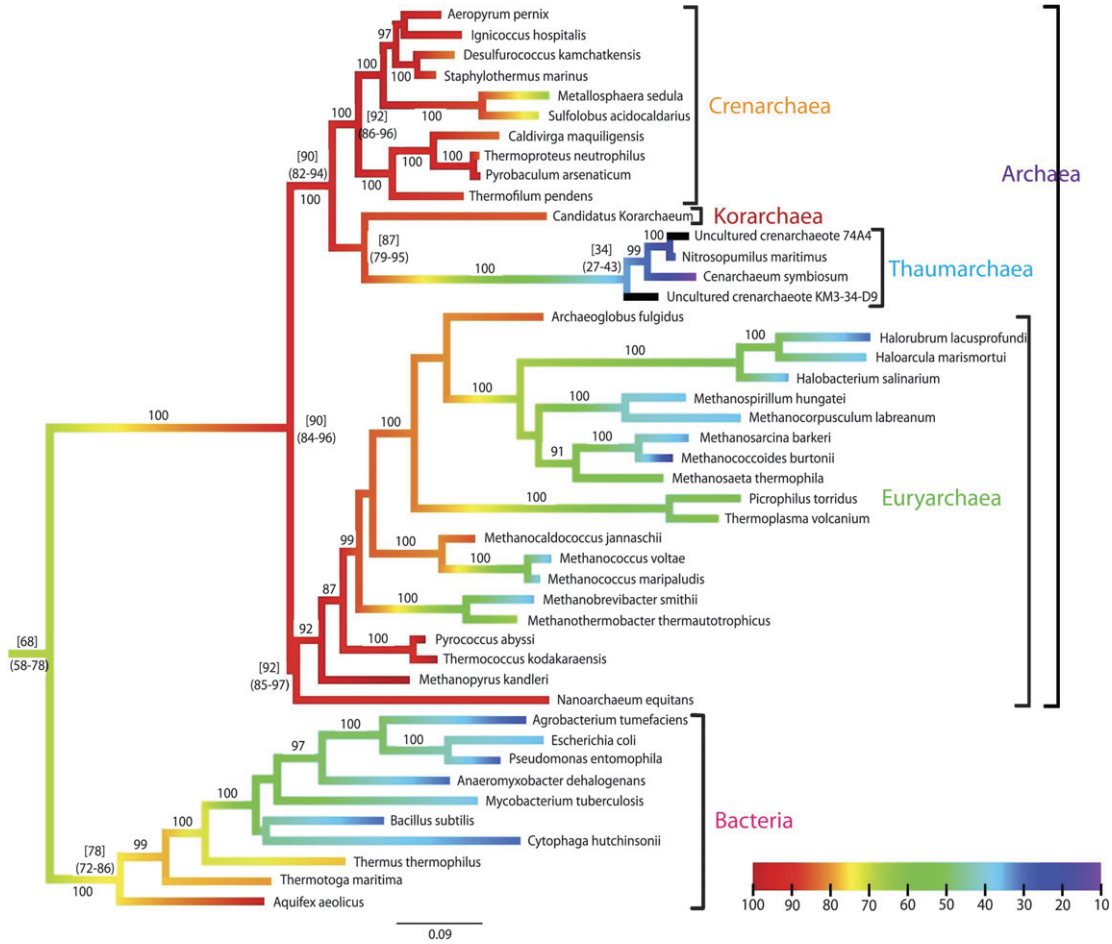



**Fig. 2.** Evolution of OGT from a hyperthermophilic ancestral state over the rRNA archaeal tree. Branch lengths have been colored according to temperature estimates at nodes. A linear gradient of color has been drawn between nodes. No evolution of OGT is represented in the vertical tree lines. As OGTs for uncultured Thaumarchaea are not available, their branches are black colored. The branch length scale is in substitution per site. The color scale is in degree Celsius. Mean estimates of temperature at key nodes are given between square brackets. Confidence intervals (95%) for estimates of ancestral OGTs are given between round brackets. Bootstrap values higher than 85% are represented. The concatenation of small and large rRNA subunits provided an alignment of 3,719 positions.

sequences were then computed. For each ancestral node, the mean of the distribution of G+C content values was determined and projected in the previously established correlation. Concerning proteins, the amino acid composition of each ancestral sequence was added to the correspondence analysis to get its projection on the second factor. Finally, the mean of the distribution of second factor values was used to infer OGT.

Figures 2 and 3 show that there is a parallel adaptation to high temperatures from a common ancestor of Archaea and Bacteria to a common ancestor of each domain. The last archaeal common ancestor is predicted to be hyperthermophilic and will be named below the HACA (Hot Archaeal Common Ancestor). From the HACA, whose OGT is estimated around 82 °C by proteins (90 °C by rRNAs), there is a slight increase of OGT until ancestors of Euryarchaea on one side (83 °C) and of Crenarchaea, Thaumarchaea, and Korarchaea on the other side (85 °C), but this increase is not statistically significant if confidence intervals are taken into account. Among Crenarchaea, OGT seems to increase along the tree until extant species such as *Aeropyrum pernix* (95 °C).

A progressive adaptation to lower temperatures is observed along the Euryarchaeal clade, similarly to what Boussau et al. (2008) observed within the bacterial domain. The euryarchaeal ancestor is predicted to be hyperthermophilic (83 °C and 92 °C for proteins and rRNAs, respectively) and deep-branching species are also adapted to these high temperatures. An adaptation of Euryarchaea to lower temperatures can then be observed with the exception of *Archaeoglobus fulgidus* and *Methanocaldococcus jannaschii* which may have readaptated to higher temperatures. The OGTs inferred for HACA are markedly higher in the present study (74–89 °C) than in the Boussau et al. (2008) study (59–73 °C). We investigated the reason(s) why the credibility intervals were not overlapping between the two studies. Three hypotheses were tested: the influence of the taxon sampling, the model of sequence evolution, and the gene sampling. We ruled out the taxon sampling and the model of sequence evolution hypotheses.

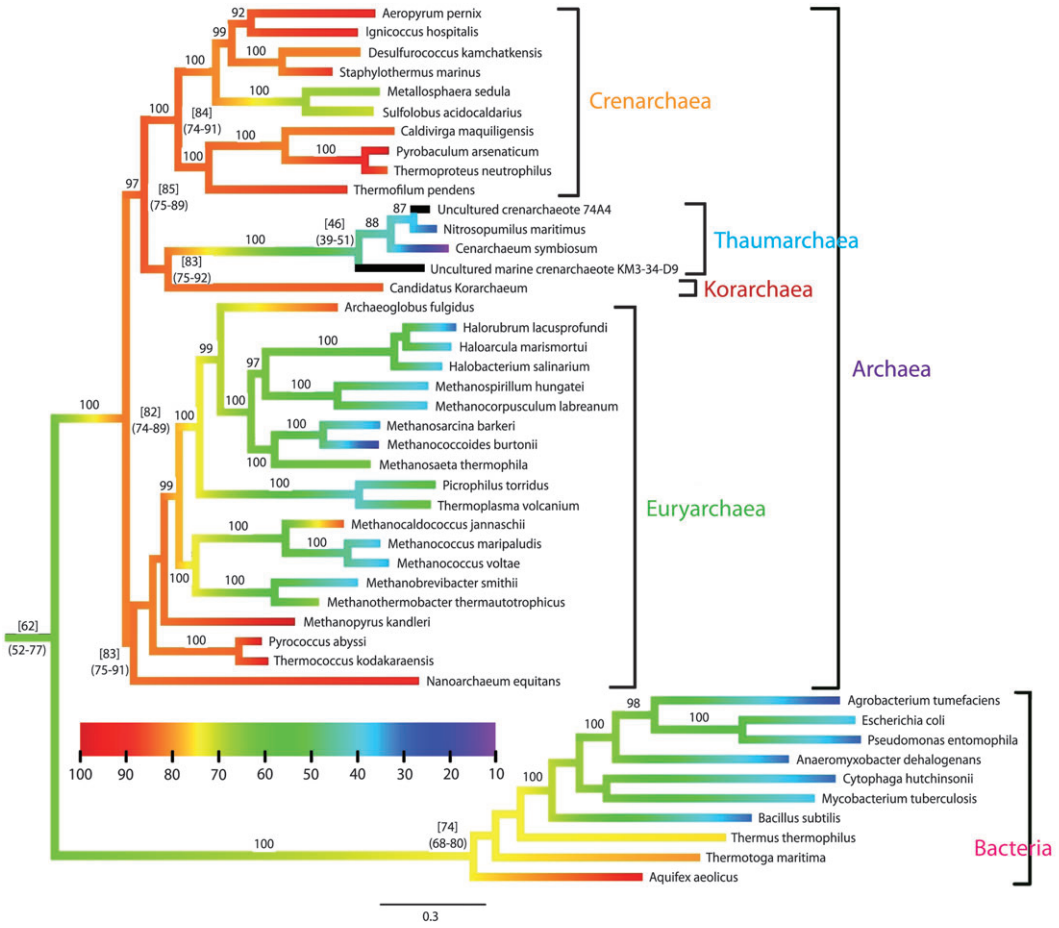



**Fig. 3.** Evolution of OGT from a hyperthermophilic ancestral state over the protein archaeal tree. See legend of figure 2 for details. The phylogenetic reconstruction is based on 72 genes and on a 9,799 amino acid long alignment.

Indeed, when the seven archaeal species used by Boussau et al. (2008) are analyzed with our data set, the ancestral OGT for HACA remains 82 °C, as with our 35 archaeal species. Furthermore, we analyzed the Boussau et al. (2008) data set with our model of sequence evolution, whereas in Boussau et al., the data were analyzed using a Bayesian approach and the CAT-BP (Blanquart and Lartillot 2008) model. The authors inferred that HACA lived around 66 °C. Here, using maximum likelihood and an a priori attribution of equilibrium frequencies along the tree to relax the homogeneity property, we obtained results very similar to Boussau et al.'s: the mean OGT inferred for HACA is 69 °C. We conclude that the difference between the two ancestral temperatures is a consequence of two interacting factors: 1) gene sampling and 2) uncertainty of the molecular thermometers. Among the 56 protein families selected by Boussau et al. (2008), only sites with less than 5% gaps were conserved. We observed that among these 56 families, only 24 contain archaeal sequences and so contributed to their final alignment. Among these 24 families, 22 are present in our 72 protein families data set. Thus, we split our data set in two parts: one with the proteins also analyzed by Boussau et al. (2008) and one with the remaining protein families. We observed that whereas the added families still predict a temperature around 83 °C, the families that are common with the Boussau et al. data set predict a temperature of 78 °C. Second, the molecular thermometer used by Boussau et al. was inferred from their own data set, using data from the bacterial domain. Here, the molecular thermometer was inferred from archaeal species only. In both approaches, the uncertainty in the regression was not taken into account and would increase the final credibility intervals.

## Evaluation of the Two Thermometers

Remarkably, rRNAs and proteins converge to quite similar estimated OGTs (figs. 2 and 3). However, the differences between rRNA and protein-based OGT predictions could result from a different signal between rRNAs and proteins. For most internal nodes, rRNAs tend to predict lower OGTs for low temperatures (<65 °C) and higher OGTs for high temperatures (>65 °C) than proteins. Thus, a negative correlation ($r = -0.79$) exists between the differences of prediction (Protein − rRNA) and the rRNA predictions (supplementary fig. 5*A*, Supplementary Material online). Interestingly, the same profile occurs with extant molecules: If the molecular thermometers are used to estimate OGTs based on the compositions of extant rRNAs and proteins, a similar negative correlation ($r = -0.53$) is found (supplementary fig. 5*B*, Supplementary Material online).

Consequently, one can reasonably assume that the differences of OGT prediction between rRNAs and proteins for internal nodes result from the differences of precision of the thermometers between each other and not from a different signal that these two molecules intrinsically carry. Finally, the precision of each thermometer was investigated. The OGTs predicted from rRNAs and proteins for extant species were compared with the reference OGTs found in the databases (supplementary fig. 6, Supplementary Material online). For both rRNAs and proteins, there is a strong positive correlation between the two variables ($r = 0.95$ and $r = 0.8$, respectively), which proves that both thermometers are reliable. However, the sum of the squares of deviations to the $y = x$ line is much lower for rRNAs than for proteins (3,298 against 11,202), which indicates that the rRNA thermometer tends to be more precise than the protein thermometer.

## The Influence of the Input Topology

The estimations of ancestral OGTs are not sensitive to the initial topology. Indeed, OGTs have been estimated at each node for the 15 topologies. The variance is small for all OGTs, and the HACA, euryarchaeal, and crenarchaeal ancestors are always predicted to be hyperthermophilic (supplementary table 4, Supplementary Material online). In general, for all ancestral nodes of each phylum, the pattern observed with the topology no. 14 stays unchanged. The same conclusion can be drawn concerning the possible influence of Eukarya. Supplementary table 5 (Supplementary Material online) shows that inferences of ancestral OGTs remain roughly the same for crucial nodes (Crenarchaea, Thaumarchaea, Korarchaea, and Euryarchaea) of the four archaeal domains of the eocyte tree. In particular, the common ancestor of all archaeal groups and Eukarya is still inferred to be hyperthermophilic.

## Temperature Is the Major Selective Pressure Governing Sequence Evolution in Archaea

Several hypotheses have been developed to explain the adaptation of mesophilic archaea to low OGTs (López-García et al. 2004). A classical scenario posits that mesophilic species have adapted from their thermophilic ancestor to cold temperatures through extensive HGTs from mesophilic bacteria or other archaea. Such a hypothesis could explain why some species were able to colonize other ecological niches and became adapted to cold environments. Recently, Drake showed that thermophilic species display very low mutation rates in comparison to mesophilic species (Drake 2009). So, once the process of adaptation to colder temperatures begun, one can propose that the lineages were subjected to a relaxation of negative selective pressures and to an increase of mutational rates. Our results strongly support this hypothesis. Indeed, there is a strong positive correlation ($r = 0.82$, $P$ value $< 0.001$) between deviations in rRNA G+C content and evolutionary distances (branch lengths) from the HACA to leaves (fig. 4A). Mesophilic species tend to have long branches associated with a strong deviation of rRNA G+C content from the HACA. Concerning proteins, the correlation between the deviation of second factor values of the correspondence analysis and branch lengths is also statistically significant, but weaker ($r = -0.57$) (data not shown). However, it has been reported that amino acid compositions of thermo- or hyperthermophiles are strongly linked to OGT. Thus, Hickey and Singer (2004) and Tekaia et al. (2002) showed that proteins of thermophilic species are slightly enriched in charged residues (Glu, Arg, Lys), whereas being depleted in polar uncharged (Asn, Gln, Ser, Thr) and in thermolabile residues (His, Gln, Thr). So, the evolution of the protein compositions for these amino acids was studied. A high correlation between deviations of ERK content (Glu, Arg, Lys) from the HACA to species and branch lengths exists ($r = 0.74$) (fig. 4C) and so does between HQT (His, Gln, Thr) content ($r = -0.55$, $P$ value $= 0.001$) or NSTQ (Asn, Gln, Ser, Thr) content ($r = -0.69$, $P$ value $< 0.001$) and branch lengths (supplementary fig. 3A and C, Supplementary Material online). As discussed above, the regressions of figure 4A and C may be potentially biased by the nonindependence of the data points. Thus, the PIC approach was also employed here to measure the statistical significance of the relationship between evolutionary rates and variations of composition. Figure 4B and D compares PICs in the evolutionary rates and in GC contents for rRNAs and ERK contents for proteins, respectively. The correlation coefficients equal 0.76 and 0.67, respectively, and are highly significant ($P$ value $< 0.001$). Furthermore, similar conclusions can be drawn for the HQT and NSTQ contents (supplementary fig. 3B and D, Supplementary Material online), with correlation coefficients of $-0.35$ ($P$ value $< 0.05$) and $-0.65$ ($P$ value $< 0.001$), respectively. Consequently, intrinsic molecular evolution of rRNAs and proteins co-occurred with the continuous adaptation of mesophilic species to colder environments.

## Control for a Putative Bias in the Nonhomogeneous Approach

In the archaeal domain, short branches have a small variation of G+C content, and long branches have a high variation. One could argue that a bias exists in our nonhomogeneous estimation of evolutionary parameters, which would systematically lead to this pattern. Of course, short branches exclude high G+C content variations between the two branch extremities, but long branches could exist with and without extensive base composition variation. Thus, a simulation experiment was carried out for the rRNA data set, where the association between branch lengths and variation of G+C content is the highest (figs. 2 and 4A). A model tree with the topology of the archaeal domain tree was used. Random branch lengths and random G+C equilibrium frequencies were attributed to each branch of this tree. Branch lengths were randomly extracted from 95% of a Poisson distribution with a mean chosen to preserve the total variance of branch lengths of the archaeal tree. The G+C equilibrium frequencies were extracted

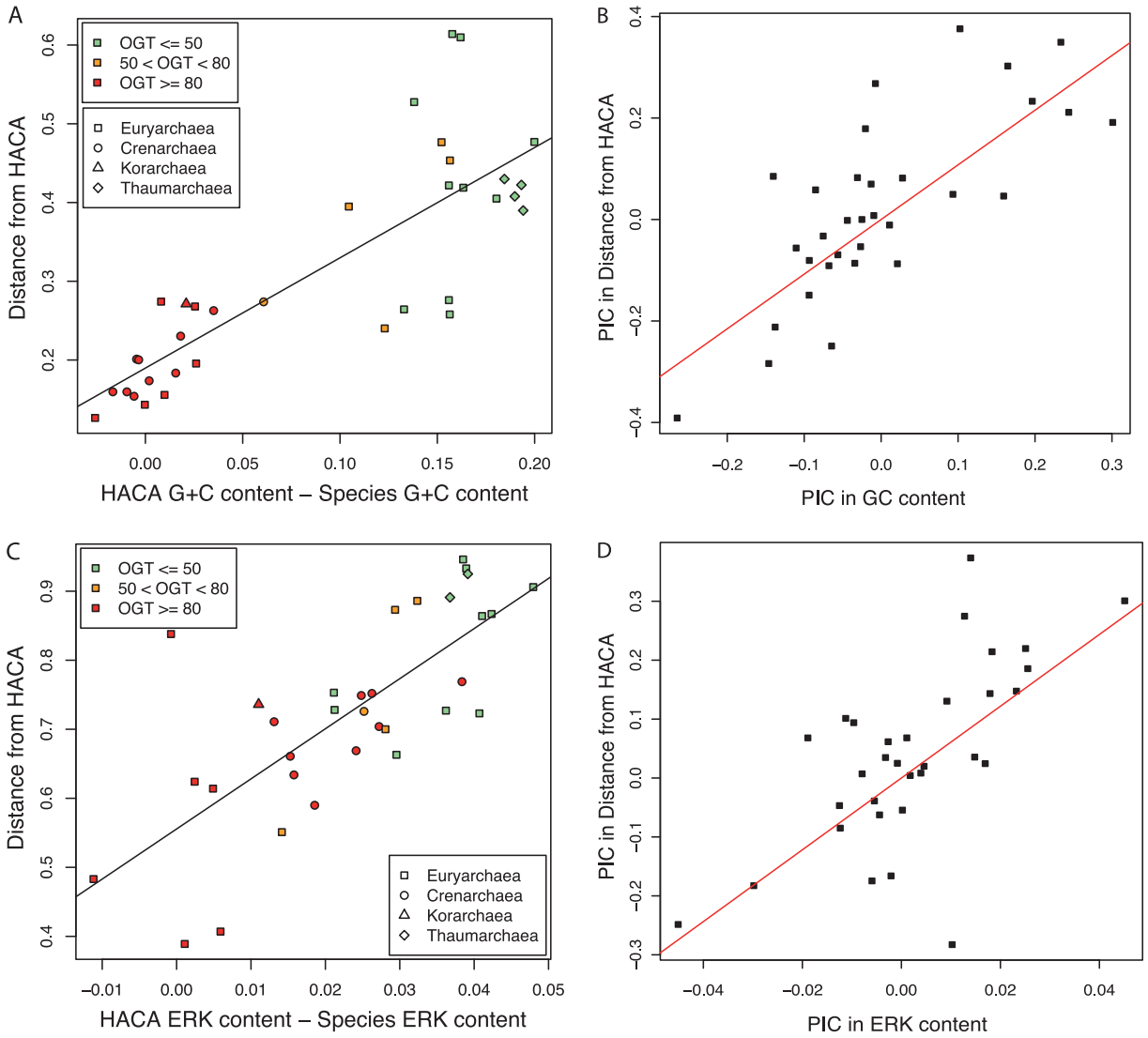



**Fig. 4.** A nonclock behavior of archaeal rRNAs and proteins and its relation with environmental temperature. In figure 4*A* and *C*, correlations between the raw evolutionary distances from HACA (total of branch lengths between the HACA and the extant species) and raw deviations of specific base (G+C%, fig. 4*A*) and amino acid (E+R+K%, fig. 4*C*) contents between the HACA and extant species are represented. E, R, and K amino acids represent charged residues. Mesophilic species are colored in green, thermophilic in orange, and hyperthermophilic in red. The four major groups of Archaea are plotted with different symbols. The linear correlation coefficient is 0.82 ($P$ value $< 0.001$) for rRNAs and 0.74 for proteins ($P$ value $< 0.001$). In figure 4*B* for rRNAs and figure 4*D* for proteins, the raw values have been corrected using the PICs method. The correlation coefficients are 0.76 ($P$ value $< 0.001$) and 0.67 ($P$ value $< 0.001$), respectively.

from 95% of a normal distribution with mean and variance equal to those of the archaeal tree. One hundred trees were constructed in this way. Each simulated tree was used to reconstruct simulated alignments with BppSeqGen, which then were used as input alignments to run the HKY85θpb model with BppML. Using simulated sequences, we reestimated the ancestral G+C content of the HACA and computed correlations between differences in G+C content and evolutionary distances from the HACA to leaves. The resulting distribution of correlation coefficients (supplementary fig. 4,Supplementary Material online) has a mean value of 0.23 and a maximum value of 0.69, far from the correlation coefficient of the real data ($r = 0.82$). This result strongly suggests that the observed pattern is a real signal and not a bias of the nonhomogeneous parameter estimation protocol.

## The Node-Density Artifact

Mesophilic species could have longer branch lengths because of the node-density artifact highlighted by Webster et al. (2003). This phenomenon could be particularly true in the Euryarchaeal domain, where more bifurcations exist, but does not apply to the long branch leading to thaumarchaeal species, which is poor in internal nodes. To rule out this possible bias in the euryarchaeal domain, eight euryarchaeal species were removed from the analysis and a nonhomogeneous experiment with the HKY85θpb model was carried out. The resulting tree (data not shown)

still displays longer branch lengths for euryarchaeal mesophilic species than for thermophilic species, which disproves the node-density artifact.

## Discussion

The main goal of this study was to investigate the evolution of the adaptation to OGT in Archaea over evolutionary times. We do not claim that the archaeal topology (figs. 2 and 3) used to infer ancestral compositions and OGTs reflects the true evolutionary history. Indeed, many protein families used here are likely to have been affected by HGT. However, the results above show that the chosen archaeal tree does not strongly influence the OGT estimates of the ancestors of the major archaeal phyla.

The nonhomogeneous models employed better fit the data than homogenous ones and allow for a more realistic description of the evolutionary process. Moreover, many archaea-specific gene families that do not have members in Bacteria were used. This further increased the signal to estimate ancestral compositions in the archaeal domain in comparison to the work reported by Boussau et al. (2008). Here, as Galtier and Gouy (1998) already highlighted with their nonhomogeneous model (one T92 substitution model per branch with $\kappa$ [transversion/transition ratio] shared in the whole tree), we confirmed that the crucial parameter which characterizes the evolution of rRNAs is $\theta$ (G+C content). However, the use of one HKY85 model per branch with $\kappa$, $\theta_1$, and $\theta_2$ shared in the whole tree, allowed to significantly improve phylogenetic estimations: the two additional parameters ($\theta_1$ and $\theta_2$) remove the constraint of equal base frequencies assumed in the T92 model and enhance the model fit to the data. The Bowker's test used in this study was implemented for DNA or RNA. It suggests that the HKY85$\theta$pb model was suitable to fit the heterogeneity of the rRNA data set. A rather similar approach that could be applied to the protein data set was proposed by Foster (2004).

The evolution of OGT along the archaeal tree presented here is in line with previous studies that used phylogenetics to infer ancestral conditions of life (Galtier et al. 1999; Boussau et al. 2008; Gaucher et al. 2008). Nevertheless, this is the first one that reaches such a level of accuracy for one particular domain of life. Our results show that the archaeal domain, from an ancestral state adapted to high temperatures, progressively colonized colder environments on Earth in the euryarchaeal phylum. This evolution of OGT is very similar to the one reconstructed for the bacterial domain (Boussau et al. 2008; Gaucher et al. 2008). In Thaumarchaea, partial genomic sequences of *Uncultured crenarchaeote 74A4* and *Uncultured crenarchaeote KM3-34-D9* were used to infer the ancestral sequences and OGTs (only one gene of each organism was present in the protein alignment). This did not cause any bias in ancestral estimations of OGTs because no major inconsistency was observed at internal nodes of this phylum in comparison with the rRNA tree.

Remarkably, the global pattern of OGT predictions is qualitatively similar between the two data sets. Even if some discrepancies exist between rRNAs and protein-based inferences, the results presented here suggest that these differences do not result from different evolutionary signals carried by rRNAs and proteins but originate from the specific prediction bias of each thermometer.

It is worthwhile to note that the difference between the present study and that by Boussau et al. (2008) concerning the OGT of HACA reveals the uncertainty in the current approach regarding the estimation of ancestral OGTs. We have ruled out that this difference in inferred OGTs resulted from the differences in taxon sampling or in evolutionary models between the two studies and have shown that the difference mostly results from different protein gene sets. In future approaches, the uncertainties of the molecular thermometers should be incorporated in the inferences of temperatures, allowing to improve the modeling of the evolution of protein sequences, without constraining it to a single dimension (here the regression line).

How did organisms that were adapted to life at high temperatures acquire the ability to colonize colder environments? An attractive hypothesis would be to bring into play intensive HGTs between archaeal species that live in hot environments and other species (bacterial or archaeal) that live in colder ones. In their analysis of partial genomic sequence data from mesophilic crenarchaea, Lopez-Garcia et al. (2004) proposed that HGTs could have been crucial in the adaptation of Thaumarchaea to cold environments. They mentioned the case of the HSP70 chaperone, present in mesophilic euryarchaea and thaumarchaea but not in hyperthermophilic crenarchaea. They supposed that the gene could have been acquired by HGT and could have facilitated the adaptation of thaumarchaea to lower temperatures. At present, with more completely sequenced genomes, this assumption remains valid because no hyperthermophilic crenarchaeon possesses this gene (data not shown). Moreover, with a deeper analysis of more newly sequenced thaumarchaeal fosmids, these authors showed that chromosomal rearrangements in the region of the rRNA genes occurred during the evolution of Thaumarchaea, more than in other lineages, and that many HGTs from bacterial and mesophilic euryarchaeal lineages can be highlighted (Brochier-Armanet C, personal communication).

How did archaeal genomes and proteomes evolve during the transition from thermophilic to mesophilic environments? Concerning rRNAs, it has been shown that thermo- or hyperthermophilic organisms display especially high values of G+C%. As rRNAs possess a large fraction of double-stranded regions, a high G+C% could provide a higher stability at high temperatures (Galtier and Lobry 1997). Concerning proteins from organisms living in hot temperatures, they are very stable from both a thermodynamic and a kinetic point of view (Sterner and Liebl 2001). Some estimations focusing on the comparison between the two bacteria *Escherichia coli* and *Thermus thermophilus*

for the RNAse H have shown that the melting temperature of the thermophilic protein is 20 °C higher than that of the mesophilic species (Hollien and Marqusee 1999). The literature mentions a lot of characters to explain why thermophilic proteins are so thermostable. Sterner and Liebl (2001) proposed a nonexhaustive list of several characters that could avoid chemical degradation of the polypeptide chain, such as an increase of hydrogen bonds, improved electrostatic interactions, and increased compactness. The results of Tekaia et al. (2002) and Hickey and Singer (2004) are good evidence to support these trends: proteins of thermophilic species are slightly enriched in charged residues (Glu, Arg, Lys), whereas being impoverished in polar uncharged (Asn, Gln, Ser, Thr) and in thermolabile residues (His, Gln, Thr).

Thus, the highly correlated evolutionary changes observed with the PIC method between branch lengths and the variation of GC (for rRNAs), ERK, NSTQ, and HQT (for proteins) contents between the HACA and extant species are informative. This phenomenon offers an explanation of the deviation from molecular clock in the archaeal domain and highlights the critical role played by environmental temperature on the archaeal molecules. An exception to this evolutionary scenario concerns *N. equitans* and its very long branch. Indeed, this organism developed a parasitic life at the surface of its hyperthermophilic crenarchaeal host, *I. hospitalis* (Hubert et al. 2002). This way of life could explain the increase of evolutionary rates unrelated to temperature in this lineage.

The use of nucleotide and amino acid sequences to estimate the timing of the history of life on earth was proposed early on (Zuckerkandl and Pauling 1965) in the history of molecular biology. The molecular clock hypothesis assumed that molecules evolved at constant rates over time, which allowed the inference of divergence times between species from molecular sequence data. However, it was later clearly demonstrated that evolutionary rates are not constant over time, either between lineages or within a particular lineage. The branches of the present archaeal domain phylogenetic trees clearly differ extensively in length, which contradicts the molecular clock model. Nowadays, relaxed molecular clock methods are developed to take these variations of evolutionary rates into account (Thorne et al. 1998; Rannala and Yang 2007).

Recent studies (Friedman et al. 2004; Drake 2009) seem to confirm that increased temperature imposes increased constraints on genetic innovation. Species adapted to high temperatures exhibit very low mutational rates. The possibility that neutral or slightly deleterious mutations in cold environments may become highly deleterious in hot temperatures, especially concerning protein folding, could explain this phenomenon. Thus, the increase of evolutionary rates during the colonization of cold environments is partly explained by increased possibilities to explore the substitutional space, without fitness impact. However, if only a neutral process of evolution were involved, we would expect to observe a broader range of G+C content and second factor values among extant mesophilic species rRNAs in figure 1*A*. All mesophilic species rRNAs are G+C poor and are characterized by low values of the second factor of the correspondence analysis. Therefore, natural selection forces archaeal organisms to have low substitutional rates in hot temperatures but, even if mesophilic species can have higher mutational and substitutional rates and are freer to explore more genetic combinations, environmental temperature continuously constrains the base and amino acid equilibrium frequencies.

In conclusion, mesophilic species have adjusted their molecular compositions during an adaptation process to colonize new mesophilic environments. The results obtained by Cherry (2010) support this view. The author showed in five eukaryotes and one mesophilic bacterium (*E. coli*) that highly expressed and slowly evolving proteins have similar compositions to those of proteins from thermophilic organisms. Because the amino acid composition of thermophilic proteins has been shown to increase the stability of the folded state at high temperatures (Singer and Hickey 2003), Cherry (2010) proposed that there is a strong selection against protein misfolding. This selection would be higher for highly expressed proteins and would be the reason of their low evolutionary rates (Pál et al. 2001; Drummond et al. 2005). In Eukaryotes (Cherry 2010) and mesophilic bacteria (Rocha and Danchin 2004) (and probably mesophilic archaea), this selection is higher for highly expressed genes. For thermophilic species (Archaea and probably Bacteria), environmental temperature appears to be a major selective factor at the whole proteome level, explaining the decrease of evolutionary rates in thermophilic proteins.

## Supplementary Material

Supplementary figures 1–6 and tables 1–5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/)

## Acknowledgments

The authors would like to thank four anonymous referees as well as the Associate Editor for their constructive comments, which allowed to significantly improve the present manuscript. The authors are particularly grateful to Céline Brochier-Armanet, Bastien Boussau, and Vincent Daubin for their help, suggestions, and fruitful discussions. Finally, the authors sincerely thank Julien Dutheil for all his help with the Bio++ libraries.

## References

Ababneh F, Jermiin LS, Ma C, Robinson J. 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. *Bioinformatics* 22:1225–1231.

Blanquart S, Lartillot N. 2008. A site- and time-heterogeneous model of amino acid replacement. *Mol Biol Evol*. 25:842–858.

Boussau B, Blanquart S, Necsulea A, Lartillot N, Gouy M. 2008. Parallel adaptations to high temperatures in the Archaean eon. *Nature* 456:942–945.

Boussau B, Gouy M. 2006. Efficient likelihood computations with nonreversible models of evolution. *Syst Biol*. 55:756–768.

Brochier C, Gribaldo S, Zivanovic Y, Confalonieri F, Forterre P. 2006. Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales? *Genome Biol*. 6:R42.

Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P. 2008. Mesophilic crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nat Rev Microbiol*. 6:245–252.

Bromham L. 2009. Why do species vary in their rate of molecular evolution? *Biol Lett*. 5:401–404.

Bromham L, Rambaut A, Harvey PH. 1996. Determinants of rate variation in mammalian DNA sequence evolution. *J Mol Evol*. 43:610–621.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*. 17:540–552.

Cherry JL. 2010. Highly expressed and slowly evolving proteins share compositional properties with thermophilic proteins. *Mol Biol Evol*. 27:735–741.

Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. 2008. The archaebacterial origin of eukaryotes. *Proc Natl Acad Sci U S A*. 105:20356–20361.

DeLong EF. 1992. Archaea in coastal marine environments. *Proc Natl Acad Sci U S A*. 89:5685–5689.

Denamur E, Matic I. 2006. Evolution of mutation rates in bacteria. *Mol Microbiol*. 60:820–827.

Drake JW. 2009. Avoiding dangerous missense: thermophiles display especially low mutation rates. *PloS Genet*. 5(6):e1000520.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A*. 102:14338–14343.

Dutheil J, Boussau B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol Biol*. 8:255.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*. 5:113.

Elkins JG, Podar M, Graham DE, et al. (20 co-authors). 2008. A korarchaeal genome reveals insights into the evolution of the Archaea. *Proc Natl Acad Sci U S A*. 105:8102–8107.

Felsenstein J. 1985. Phylogenies and the comparative method. *Am Nat*. 125:1–15.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Forterre P, Gribaldo S, Brochier-Armanet C. 2009. Happy together: genomic insights into the unique Nanoarchaeum/Ignicoccus association. *J Biol*. 8:7.

Foster PG. 2004. Modeling compositional heterogeneity. *Syst Biol*. 53:485–495.

Foster PG, Cox CJ, Embley TM. 2009. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. *Philos Trans R Soc Lond B Biol Sci*. 364:2197–2207.

Foster PL. 2007. Stress-induced mutagenesis in bacteria. *Crit Rev Biochem Mol Biol*. 42:373–397.

Friedman R, Drake JW, Hughes AL. 2004. Genome-wide patterns of nucleotide substitution reveal stringent functional constraints on the protein sequences of thermophiles. *Genetics* 167:1507–1512.

Galtier N, Gouy M. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol Biol Evol*. 15:871–879.

Galtier N, Lobry JR. 1997. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J Mol Evol*. 44:632–636.

Galtier N, Tourasse N, Gouy M. 1999. A nonhyperthermophilic common ancestor to extant life forms. *Science* 283:220–221.

Gaucher EA, Govindarajan S, Ganesh OK. 2008. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* 451:704–707.

Gowri-Shankar V, Rattray M. 2006. On the correlation between composition and site-specific evolutionary rate: implications for phylogenetic inference. *Mol Biol Evol*. 23:352–364.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol*. 27:221–224.

Gribaldo S, Brochier-Armanet C. 2006. The origin and evolution of Archaea: a state of the art. *Philos Trans R Soc Lond B Biol Sci*. 361:1007–1022.

Grogan DW, Carver GT, Drake JW. 2001. Genetic fidelity under harsh conditions: analysis of spontaneous mutation in the thermoacidophilic archaeon Sulfolobus acidocaldarius. *Proc Natl Acad Sci U S A*. 98:7928–7933.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 52:696–704.

Harvey PH, Pagel MD. 1991. The comparative method in evolutionary biology. Oxford: Oxford University Press.

He Q, Sanford RA. 2003. Characterization of Fe(III) reduction by chlororespiring *Anaeromyxobacter dehalogenans*. *Appl. Environ. Microbiol*. 69:2712–2718.

Hegan PS, Mermall V, Tilney LG, Mooseker MS. 2007. Roles for *Drosophila melanogaster* Myosin IB in maintenance of enterocyte brush-border structure and resistance to the bacterial pathogen *Pseudomonas entomophila*. *Mol Biol Cell*. 18:4625–4636.

Hickey D, Singer G. 2004. Genomic and proteomic adaptations to growth at high temperature. *Genome Biol*. 5:117.1–117.7.

Hollien J, Marqusee S. 1999. A thermodynamic comparison of mesophilic and thermophilic ribonucleases H. *Biochemistry* 38:3831–3836.

Hubert H, Hohn MJ, Rachel R, Fuchs T, Wimmer VC, Stetter KO. 2002. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* 417:63–67.

Jobb G, von Haeseler A, Strimmer K. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol Biol*. 4:18.

Lanfear R, Welch JJ, Bromham L. 2010. Watching the clock: studying variation in rates of molecular evolution between species. *Trends Ecol Evol*. 25:395–503.

Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol*. 7:S4.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol*. 21:1095–1109.

Lake JA, Henderson E, Oakes M, Clark MW. 1984. Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proc Natl Acad Sci U S A*. 81:3786–3790.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. *Mol Biol Evol*. 25:1307–1320.

Le SQ, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24:2317–2323.

Lopez-Garcia P, Brochier C, Moreira D, Rodriguez-Valera F. 2004. Comparative analysis of a genome fragment of an uncultivated mesopelagic crenarchaeote reveals multiple horizontal gene transfers. *Environ Microbiol*. 6:19–34.

Mackwan RR, Carver GT, Drake JW, Grogan DW. 2007. An unusual pattern of spontaneous mutations recovered in the halophilic archaeon *Haloferax volcanii*. *Genetics* 176:697–702.

Mackwan RR, Carver GT, Kissling GE, Drake JW, Grogan DW. 2008. The rate and character of spontaneous mutation in *Thermus thermophilus*. *Genetics* 180:17–25.

Nazar RN. 1980. A 5.8 S rRNA-like sequence in prokaryotic 23 S rRNA. *FEBS Lett*. 119:212–214.
Pál C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. *Genetics* 158:927–931.
Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290.
Penel S, Arigon AM, Dufayard JF, Sertier AS, Daubin V, Duret L, Gouy M, Perrière G. 2009. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*. 10:S3.
Rannala B, Yang Z. 2007. Inferring speciation times under an episodic molecular clock. *Syst Biol*. 56:453–466.
Ripplinger J, Sullivan J. 2008. Does choice in model selection affect maximum likelihood analysis? *Syst Biol*. 57:76–85.
Rocha EP, Danchin A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol Biol Evol*. 21:108–116.
Roure B, Rodriguez-Ezpeleta N, Philippe H. 2007. SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evol Biol*. 7(Suppl 1):S2.
Singer G, Hickey D. 2003. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene* 317:39–47.
Sterner R, Liebl W. 2001. Thermophilic adaptation of proteins. *Crit Rev Biochem Mol Biol*. 36:39–106.
Stetter KO. 2006. Hyperthermophiles in the history of life. *Philos Trans R Soc B Biol Sci*. 361:1837–1843.
Strimmer K, Rambaut A. 2002. Inferring confidence sets of possibly misspecified gene trees. *Proc R Soc B Biol Sci*. 269:137–142.
Tekaia F, Yeramian E, Dujon B. 2002. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. *Gene* 297:51–60.
Thioulouse J, Chessel D, Dolédec S, Olivier J. 1997. ADE-4: a multivariate analysis and graphical display software. *Stat Comput*. 7:75–83.
Thompson JD, Gibson TJ, Higgins DG. 2002. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics*. Chapter 2:Unit 2.3.
Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol*. 15:1647–1657.
Valentine DL. 2007. Adaptations to energy stress dictate the ecology and evolution of the Archaea. *Nat Rev Microbiol*. 5:316–323.
Vetriani C, Maeder DL, Tolliday N, Yip KS, Stillman TJ, Britton KL, Rice DW, Klump HH, Robb FT. 1998. Protein thermostability above 100 degreesC: a key role for ionic interactions. *Proc Natl Acad Sci U S A*. 95:12300–12305.
Webster AJ, Payne RJ, Pagel M. 2003. Molecular phylogenies link rates of evolution and speciation. *Science* 301:478.
Zuckerkandl E, Pauling L. 1965. Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ, editors. Evolving genes and proteins. New York: Academic Press. p. 97–166.

## 3.2 Le dernier ancêtre commun universel et ses compositions moléculaires mésophiliques.

### 3.2.1 Introduction

Cette partie traite toujours de l'adapation à la température des premières lignées de l'arbre de la vie mais cette fois-ci plus spécifiquement de la température à laquelle était adapté le dernier ancêtre commun universel (LUCA). Cette question a été le sujet de nombreux débats ces dernières années et contrairement à ce qui était classiquement pensé, des analyses basées sur l'analyse des compositions moléculaires ancestrales ont soutenu la conclusion d'un LUCA vivant à basse température (Galtier et al., 1999; Boussau et al., 2008), à partir duquel les lignées menant aux ancêtres des Bactéries et des Archées se sont adaptées à des environnements plus chauds. Cependant, la nature du signal phylogénétique réellement capturé par les modèles de substitutions utilisés pour reconstruire ces compositions n'a pas été caractérisé. Dans l'article qui suit, nous montrons qu'un lien très particulier entre variations de taux d'évolution entre sites et de composition moléculaire dans le temps a permis d'enregistrer le signal expliquant ce scénario non-parcimonieux d'adaptation à la température.

Cet article a été accepté dans le journal *Biology Letters*.

### 3.2.2 Manuscrit

biology letters

rsbl.royalsocietypublishing.org

Research

**Cite this article:** Groussin M, Boussau B, Charles S, Blanquart S, Gouy M. 2013 The molecular signal for the adaptation to cold temperature during early life on Earth. Biol Lett 9: 20130608.
http://dx.doi.org/10.1098/rsbl.2013.0608

Received: 2 July 2013
Accepted: 27 August 2013

**Subject Areas:**
evolution, bioinformatics, ecology

**Keywords:**
non-homogeneous substitution model, ancestral sequence reconstruction, optimal growth temperature, last universal common ancestor, early Earth

**Author for correspondence:**
Mathieu Groussin
e-mail: mathieu.groussin@univ-lyon1.fr

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsbl.2013.0608 or via http://rsbl.royalsocietypublishing.org.

Royal Society Publishing
Informing the science of the future

**Phylogeny**

# The molecular signal for the adaptation to cold temperature during early life on Earth

Mathieu Groussin[1], Bastien Boussau[1,2], Sandrine Charles[1], Samuel Blanquart[3] and Manolo Gouy[1]

[1]Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, Université Lyon 1, CNRS, UMR5558, Villeurbanne, France
[2]Department of Integrative Biology, University of California, Berkeley, CA, USA
[3]Inria Lille Nord Europe, LIFL UMR 8022 (CNRS Université de Lille 1), Villeneuve d'Ascq, France

Several lines of evidence such as the basal location of thermophilic lineages in large-scale phylogenetic trees and the ancestral sequence reconstruction of single enzymes or large protein concatenations support the conclusion that the ancestors of the bacterial and archaeal domains were thermophilic organisms which were adapted to hot environments during the early stages of the Earth. A parsimonious reasoning would therefore suggest that the last universal common ancestor (LUCA) was also thermophilic. Various authors have used branch-wise non-homogeneous evolutionary models that better capture the variation of molecular compositions among lineages to accurately reconstruct the ancestral G + C contents of ribosomal RNAs and the ancestral amino acid composition of highly conserved proteins. They confirmed the thermophilic nature of the ancestors of Bacteria and Archaea but concluded that LUCA, their last common ancestor, was a mesophilic organism having a moderate optimal growth temperature. In this letter, we investigate the unknown nature of the phylogenetic signal that informs ancestral sequence reconstruction to support this non-parsimonious scenario. We find that rate variation across sites of molecular sequences provides information at different time scales by recording the oldest adaptation to temperature in slow-evolving regions and subsequent adaptations in fast-evolving ones.

## 1. Introduction

Several lines of evidence support the hypothesis that, during early stages of the evolution, life was adapted to high temperatures that may have prevailed on the surface of the early Earth. For instance, previous studies discovered that the deepest branching lineages within the bacterial and archaeal domains are thermophilic [1]. This scenario is also supported by the reconstruction and synthesis of ancestral translation elongation factor Tu sequences that appear more and more thermostable when going back in time [2] and by an estimation of the amino acid composition of ancestral proteomes that appear more similar to the composition of extant thermophiles than that of mesophiles [3].

A tight relation exists between either the G + C content in ribosomal RNAs or the amino acid contents in proteins and the optimal growth temperature (OGT) of Bacteria and Archaea. Such correlations between molecular composition and temperature may be explained by structural adaptation increasing RNA and protein thermostability [4,5] and are likely to remain constant over evolutionary time. They allow the construction of molecular thermometers [6] that can provide estimates of ancestral environmental temperatures if one obtains

ancestral base and amino acid compositions through ancestral sequence reconstruction. Using such an approach, Boussau *et al*. [7] concluded that molecular sequence data confirm the hypothesis of high-temperature adaptation during the early stages of life, namely for the ancestors of the bacterial and archaeal domains. However, these authors reported strong evidence for a non-parsimonious scenario in which the last universal common ancestor (LUCA) itself, living at a still earlier stage of the history of life, was a mesophilic organism.

Through a number of control experiments, Boussau *et al*. [7] have shown that the use of non-homogeneous substitution models, which are capable of capturing the variation of composition among lineages, are key to accurately estimate ancestral base and amino acid compositions, and therefore ancestral temperatures. But these authors have not identified the specific molecular properties present in extant sequences that inform non-homogeneous models to support such a non-parsimonious scenario. In this letter, we aim to address this issue.

## 2. Material and methods

### (a) Datasets and non-homogeneous models

Boussau *et al*. [7] built a concatenate of small- and large-subunit rRNAs from 456 organisms (2239 sites) and used the sites restricted to stem regions (1043 sites) to infer the ancestral G + C contents over the tree of life. From these alignments, we selected 125 species covering a broad taxonomic diversity without redundancy in the taxonomic sampling. Regarding the concatenation of proteins, the 56 gene families and 30 species considered in Boussau *et al*. [7] were used here, and increased to 38 species, with the addition of Archaea species in particular, which were poorly represented in the first set of species. We reconstructed ML phylogenetic trees for rRNAs (on the 2239 sites dataset) and proteins with PhyML [8]. A three-domain tree was obtained and the root was placed on the branch between the ancestors of Bacteria and Archaea/Eukaryotes. As in [9] and [7], the branch-wise equilibrium frequencies were estimated along these universal phylogenetic trees. The stem dataset was analysed with the BppML program [10] assuming a discrete gamma distribution with eight categories to model rate variation among sites and the non-homogeneous Galtier & Gouy (GG) substitution model [11]. The GG model specifies branch-wise equilibrium G + C contents, as well as an independent G + C content at the root. For proteins, we used a new branch-wise non-homogeneous model implemented in the maximum-likelihood (ML) framework, named COaLA [12] that we recently designed. See the electronic supplementary material for a description of the COaLA model and an evaluation of the fit to data of the non-homogeneous models in comparison with homogeneous models.

### (b) Molecular thermometers

OGT highly correlates with the G + C content of the stem regions of rRNAs ($\rho = 0.76$, $p$-value $< 0.001$; see the electronic supplementary material, figure S2) and with the second axis of the COA computed on amino acid compositions of the protein dataset restricted to prokaryotic species ($\rho = 0.88$, $p$-value $< 0.001$; see the electronic supplementary material, figure S3). We controlled for phylogenetic inertia with the phylogenetic independent contrast approach [13] using the R package APE [14] and observed that those correlations were still strongly significant. Linear regressions between OGTs and compositions were then computed to obtain the molecular thermometers.

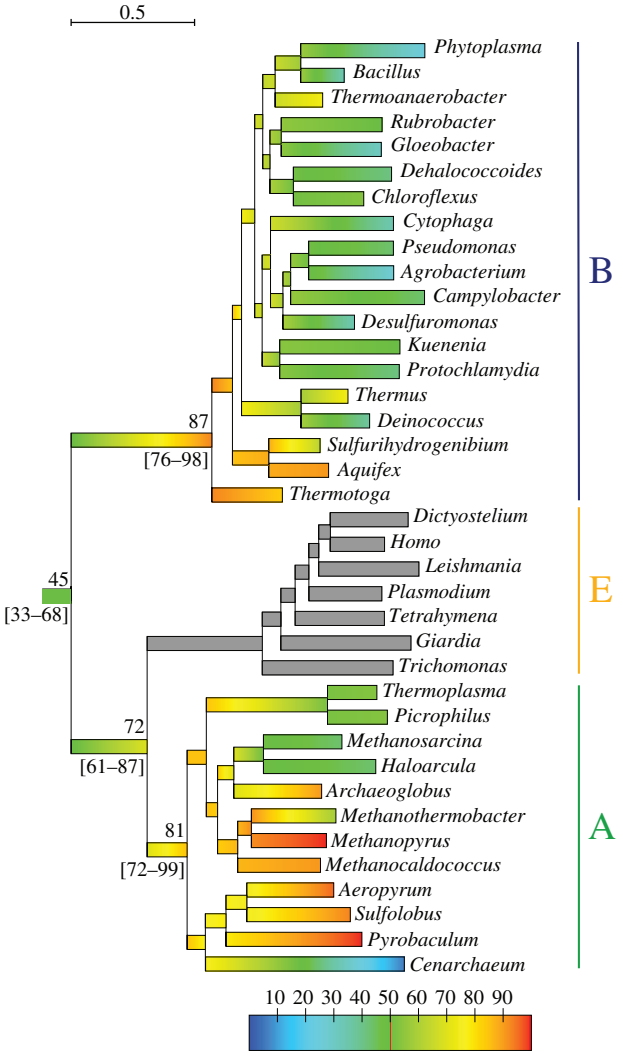



**Figure 1.** Evolution of OGT along the universal tree of life obtained with the protein dataset. Branches have been coloured according to temperature estimates at nodes, following a linear interpolation from node to node. OGTs for Eukaryotes are not available, their branches are therefore grey coloured. The branch length scale is in substitution/site. The colour scale is in °C. Mean estimates of temperature for LUCA and the ancestors of major domains are given above branches. Confidence intervals (95%) for estimates of ancestral OGTs are given between square brackets.

### (c) Inference of ancestral compositions and optimal growth temperatures

The ancestral sequences were inferred with BppAncestor [10] using the evolutionary parameters estimated by BppML. For each node of the tree, 100 ancestral sequences were generated by drawing amino acids from the posterior distributions of probabilities. The average composition of these ancestral sequences was calculated and the corresponding ancestral temperatures were deduced from the molecular thermometers (see the electronic supplementary material for the confidence intervals computation and the caution required when interpreting ancestral temperatures).

## 3. Results and discussion

We first confirm results obtained in [7] with the present rRNA and protein datasets and the non-homogeneous GG [11] and COaLA [12] substitution models in ML. Figures 1 and 2

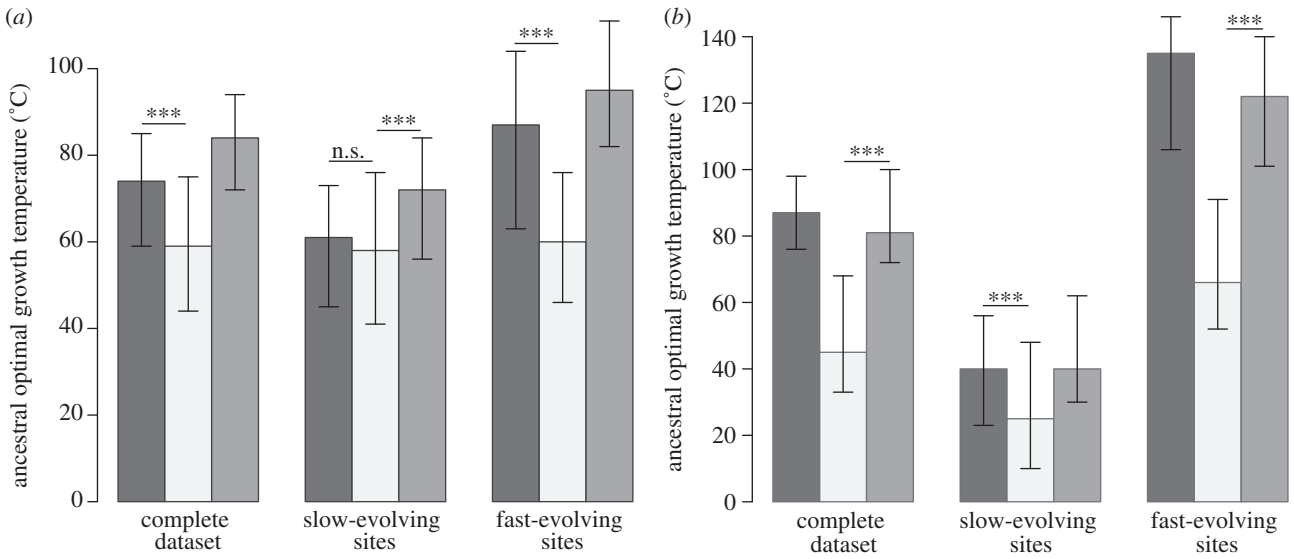



**Figure 2.** The non-homogeneous models recover the signal for a parallel adaptation to high temperatures within the across-site rate variation. (*a*) rRNA dataset. (*b*) Protein dataset. Ancestral temperatures for domain ancestors and for LUCA were estimated from ancestral compositions inferred with non-homogeneous models, either on all sites of the datasets (complete dataset) or on slow-evolving or fast-evolving sites only. ***$p$-value $< 0.001$. n.s. non-significant. Black bars, Bacteria; light grey bars, LUCA; dark grey bars, Archaea.

show that LUCA is estimated to have lived in colder environments than the ancestors of Bacteria and Archaea (Wilcoxon test, $p$-value $< 0.001$), which were hyperthermophiles. Electronic supplementary material, figure S1 shows that this pattern is also recovered when an alternative tree topology is used, in which Eukaryotes branch within Archaea (Eocyte hypothesis [15]) but is less pronounced with the homogeneous LG model, which infers a thermophilic LUCA.

The phylogenetic signal that informs a non-hyperthermophilic LUCA and yet two hyperthermophilic descendants is currently unknown. However, several points suggest that the variation in evolutionary rate among sites plays a role. First, Fournier & Gogarten [16] highlighted that amino acids that are found in higher proportions in hyperthermophilic species are rarer at slow-evolving sites. Such amino acids notably include charged residues [6]. Second, the signal for a parallel adaptation to high temperatures is partially lost when COaLA is employed without a gamma distribution to model the variation in rate among sites (see the electronic supplementary material, figure S1).

To highlight the influence of rate variation among sites in the differential recording of ancestral compositions, we partitioned the rRNA and protein datasets according to the site evolutionary rates. Figure 2*a*,*b* shows that with slow-evolving sites, all ancestors are inferred to be mesophilic organisms, LUCA being adapted to lower temperatures than its two descendants. The ancestral compositions of fast-evolving sites tend to favour hotter ancestral environments, even for LUCA with proteins. But LUCA is still inferred to live at lower temperatures than the ancestors of Bacteria and Archaea. As expected, the quantitative estimates of past temperatures inferred by both slow- and fast-evolving sites are different from those obtained with the complete dataset. Indeed, although slow-evolving sites conserved reliable signals for ancestral compositions, they carry less phylogenetic information for the early parallel adaptation to high temperature, which explains why this pattern is less pronounced than that with the complete dataset. However, both the G + C content (rRNAs) and the third axis of a correspondence analysis (proteins) computed from the slow-evolving sites of extant sequences correlate with OGT ($\rho = 0.72$, $p$-value $< 0.001$ and $\rho = 0.43$, $p$-value $< 0.05$, respectively), adding support to the idea that slow-evolving sites can respond to temperature and can represent accurate fossils of ancestral adaptation to temperature. Fast-evolving sites contain a stronger signal for this parallel adaptation but necessarily less reliable information for ancestral compositions, and so ancestral temperatures.

All these results suggest the presence of a genuine signal in molecular sequences indicating a mesophilic LUCA. This signal was recorded thanks to a combination of compositional variation in time and rate variation in site such that slow-evolving sites more accurately reflect older temperatures, while fast-evolving sites partially erased this oldest signal in favour of subsequent adaptations to higher temperatures.

Gowri-Shankar & Rattray [17] showed that there is an intrinsic correlation between evolutionary rates across sites and base composition in rRNAs. Therefore, nucleotide composition varies across the sites of an rRNA alignment. These authors showed that branch-wise non-homogeneous models, which account for the variation of composition in time but assume across-site homogeneity, may infer biased ancestral sequence compositions for sequences generated by a time-homogeneous process in which evolutionary rate and base compositions are correlated. The inference bias is directed towards the composition of slow-evolving sites which are, in the case of full-length rRNAs including both stem and loop regions, GC-poor. One could therefore wonder whether such an inference bias would be responsible for the low G + C content inferred for LUCA compared with the higher G + C contents of its first descendants. We reject this bias with two points. First, as in this study, Boussau *et al*. [7] applied the molecular thermometers on rRNAs to only the stem regions of the molecule. Electronic supplementary material, figure S4 shows that, for these regions, the correlation found by Gowri-Shankar & Rattray [17] is in the opposite direction, although non-significant, with G + C-enriched slow-evolving sites. Second, we simulated data in a context where the bias would apply, assuming only

heterogeneity among sites and no heterogeneity among branches, and verified whether the correlation between site evolutionary rates and site compositions incorrectly informs the non-homogeneous model to estimate a lower G + C content of LUCA than for its descendants. We partitioned rRNA alignment sites in eight categories according to their evolutionary rate. For each rate-specific category, we simulated DNA sequences with a homogeneous Tamura92 model and the G + C equilibrium frequency fixed to the observed G + C frequency of the category, and then concatenated the eight simulated sets. We repeated this procedure 100 times and reconstructed ancestral G + C contents with the non-homogeneous GG model on each concatenated simulated alignment. Electronic supplementary material, figure S5 shows that the pattern of parallel increase in G + C content from LUCA found from real data is not recovered. Instead, LUCA has a higher G + C content than its two descendants. As slow-evolving sites of stem regions have globally higher G + C contents than fast-evolving ones (see the electronic supplementary material, figure S4), this simulation result is in agreement with the bias of Gowri-Shankar & Rattray [17]. It further suggests that, if the non-homogeneous model applied to real data is affected by the bias as it is when applied to simulations, the true G + C content of LUCA may so far have been overestimated.

All these results indicate that non-homogeneous models can capture a genuine timewise variation in composition and that the pattern of parallel increase to high temperatures does not result from a bias owing to a correlation between site-specific rates and site-specific compositions [17] but emerges in spite of this bias.

Acknowledgements. The authors thank Nicolas Lartillot, Vincent Daubin, Chloé Tessereau, Blaise Tymen, Pierre Lévy, Florent Mazel and the members of the Bioinformatics and Evolutionary Genomics team for suggestions and fruitful discussions. The authors are also grateful to two anonymous reviewers and the editor who helped in improving this manuscript.

Data accessibility. Data are available at dryad digital repository (http://datadryad.org/): doi:10.5061/dryad.90525.

Funding statement. This work was supported by the French Agence Nationale de la Recherche (ANR) and is a contribution to the Ancestrome project (ANR-10-BINF-01-01).

## References

1. Stetter KO. 2006 Hyperthermophiles in the history of life. *Phil. Trans. R. Soc. B* **361**, 1837–1843. (doi:10.1098/rstb.2006.1907)
2. Gaucher EA, Govindarajan S, Ganesh OK. 2008 Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* **451**, 704–707. (doi:10.1038/nature06510)
3. Brooks DJ, Fresco JR, Singh M. 2004 A novel method for estimating ancestral amino acid composition and its application to proteins of the last universal ancestor. *Bioinformatics* **20**, 2251–2257. (doi:10.1093/bioinformatics/bth235)
4. Galtier N, Lobry JR. 1997 Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J. Mol. Evol.* **44**, 632–636. (doi:10.1007/PL00006186)
5. Zeldovich KB, Berezovsky IN, Shakhnovich EI. 2007 Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput. Biol.* **3**, e5. (doi:10.1371/journal.pcbi.0030005)
6. Groussin M, Gouy M. 2011 Adaptation to environmental temperature is a major determinant of molecular evolutionary rates in Archaea. *Mol. Biol. Evol.* **28**, 2661–2674. (doi:10.1093/molbev/msr098)
7. Boussau B, Blanquart S, Necsulea A, Lartillot N, Gouy M. 2008 Parallel adaptation to high temperature in the Archaean Eon. *Nature* **456**, 942–945. (doi:10.1038/nature07393)
8. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010 New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321. (doi:10.1093/sysbio/syq010)
9. Galtier N, Tourasse N, Gouy M. 1999 A nonhyperthermophilic common ancestor to extant life forms. *Science* **283**, 220–221. (doi:10.1126/science.283.5399.220)
10. Dutheil J, Boussau B. 2008 Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol. Biol.* **8**, 255. (doi:10.1186/1471-2148-8-255)
11. Galtier N, Gouy M. 1998 Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.* **15**, 871–879. (doi:10.1093/oxfordjournals.molbev.a025991)
12. Groussin M, Boussau B, Gouy M. 2013 A branch-heterogeneous model of protein evolution for efficient inference of ancestral sequences. *Syst. Biol.* **62**, 523–538. (doi:10.1093/sysbio/syt016)
13. Felsenstein J. 1985 Phylogenies and the comparative method. *Am. Nat.* **125**, 1–15. (doi:10.1086/284325)
14. Paradis E, Claude J, Strimmer K. 2004 APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290. (doi:10.1093/bioinformatics/btg412)
15. Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. 2008 The archaebacterial origin of eukaryotes. *Proc. Natl Acad. Sci. USA* **105**, 20 356–20 361. (doi:10.1073/pnas.0810647105)
16. Fournier GP, Gogarten JP. 2007 Signature of a primitive genetic code in ancient protein lineages. *J. Mol. Evol.* **65**, 425–436. (doi:10.1007/s00239-007-9024-x)
17. Gowri-Shankar V, Rattray M. 2006 On the correlation between composition and site-specific evolutionary rate: implications for phylogenetic inference. *Mol. Biol. Evol.* **23**, 352–364. (doi:10.1093/molbev/msj040)

## 3.3 Vers une systématique améliorée du monde Procaryote.

### 3.3.1 Introduction

Ce chapitre aborde deux aspects majeurs de la reconstruction phylogénétique, à savoir l'intérêt potentiel de l'utilisation de protéines ribosomiques pour résoudre les relations phylogénétiques profondes chez les Procaryotes et l'utilisation de modèles non-homogènes dans le temps afin de raciner des arbres phylogénétiques.

La taxonomie est un champ de recherche permettant de décrire les espèces vivantes et de les regrouper en entités, nommées taxons, afin de les identifier et de comprendre la structuration du vivant. L'approche phylogénétique ou systématique permet ensuite d'organiser ces entités et d'établir leurs relations entre elles. Concernant la taxonomie des Bactéries, la revue IJSEM (International Journal of Systematic and Evolutionary Microbiology) ainsi que le manuel de Bergey, dont la première édition a été écrite en 1923 par David H. Bergey, Francis C. Harrison, Robert S. Breed, Bernard W. Hammer, and Frank M. Huntoon, à la demande de la société des Bacteriologistes Americains (Society of American Bacteriologists), font souvent référence. Ces deux publications décrivent l'ensemble des caractéristiques des différents niveaux taxonomiques des Procaryotes, ainsi que des descriptions concernant la biodiversité et les milieux de culture des procaryotes. Dans les éditions les plus récentes, la taxonomie de Bergey se base en très grande partie sur des reconstructions phylogénétiques moléculaires réalisées à partir de l'ARN ribosomique 16S. Depuis les années 1970 et les travaux de Carl Woese (Woese and Fox, 1977; Fox et al., 1977), ce marqueur a été considéré comme optimal en phylogénie des Procaryotes, car il est universel, sa fonction est conservée à travers le vivant et possède des régions évoluant à la fois rapidement et lentement, permettant d'estimer à la fois des relations récentes et anciennes. Cependant, la reconstruction d'histoires phylogénétiques chez les Bactéries et les Archées représente une tâche très difficile. De nombreux phénomènes biologiques tels que les biais compositionnels, l'hétérotachie ou les transferts horizontaux peuvent empêcher de reconstruire la vraie histoire évolutive verticale des organismes. Bien que l'ARN 16S soit affecté par les biais compositionnels ou de variation de taux d'évolution dans le temps, il est moins affecté par les transferts horizontaux en comparaison avec les gènes codant des protéines. Avec l'augmentation de l'échantillonnage taxonomique permis par l'explosion actuelle des projets de séquençage de génomes complets ainsi que l'augmentation de la puissance de calcul permettant de reconstruire de grandes phylogénies, il a été remarqué que l'ARN 16S ne pouvait à lui seul permettre de reconstruire les relations de parentés entre lignées bactériennes, notamment les plus profondes. Cela démontre la nécessité de rechercher de nouveaux marqueurs pour la reconstruction phylogénétique des Prokaryotes.

Le racinement d'un arbre phylogénétique peut se réaliser à l'aide de différentes méthodes.

(i) Il se fait classiquement à l'aide d'une espèce ou groupes d'espèces externes (outgroup). Ce concept requiert l'analyse d'espèce(s) dont on sait *a priori* qu'elles n'appartiennent pas au groupe d'espèces d'intérêt (ou ingroup). Pour ce faire, les caractères homologues comparés lors de la reconstruction phylogénétique doivent incorporer ceux du groupe externe. Comme ce groupe est, par définition, externe, la divergence entre le groupe externe et le groupe interne a nécessairement pré-daté toutes les divergences du groupe interne, de telle sorte que la racine de l'arbre se positionne sur la branche reliant le groupe externe au groupe interne. Néanmoins, cette approche peut souffrir de plusieurs problèmes. Le premier est qu'il n'est pas évident de déterminer un groupe externe au groupe interne d'intérêt. L'exemple extrême est la reconstruction de l'arbre de la vie, pour lequel aucun outgroup n'est disponible. Il faut toutefois mentionner qu'il a été proposé de contourner ce problème en utilisant des gènes dupliqués ancestralement, déjà présents chez LUCA. Ainsi, il devient possible de raciner l'arbre de la vie correspondant à un des paralogues à l'aide de l'autre paralogue (Iwabe et al., 1989; Gogarten et al., 1989). Cependant, il a été montré que les marqueurs classiquement utilisés pour ces analyses (ATPase, tRNA synthetase, signal recognition particle protein, etc) n'étaient pas fiables, du fait d'une trop grande saturation mutationelle ayant éliminé la majorité du signal phylogénétique (Philippe and Forterre, 1999; Lopez et al., 1999). Le deuxième problème est que le groupe externe est souvent éloigné du groupe d'intérêt, augmentant le risque de rencontrer des problèmes d'attraction des longues branches entre certaines lignées du groupe interne et le groupe externe (Philippe et al., 2009). (ii) La seconde approche de racinement possible est l'utilisation de l'hypothèse d'horloge moléculaire : si l'on suppose que toutes les lignées évoluent à la même vitesse, le barycentre de l'arbre peut être considéré comme la racine de l'arbre (Zuckerkandl and Pauling, 1965; Kumar, 2005). Dorénavant, cette approche n'est plus considérée car les différentes analyses phylogénétiques menées sur l'ensemble des groupes du vivant ont montré à quel point l'hypothèse de constance des taux dans le temps est contredite. Cela a motivé le développement de modèles d'horloge relâchée (voir Annexes), permettant de modéliser la variation des taux de substitution dans le temps et à la fois d'estimer plus précisément les événements de divergence (Thorne et al., 1998; Rannala and Yang, 2007) et de raciner les arbres (Drummond et al., 2006). Enfin, comme les modèles non-stationnaires attribuent un ensemble de fréquences de bases ou d'acides aminés à la racine de l'arbre qui est indépendant des fréquences le long de l'arbre (voir Introduction), la position de la racine influence la vraisemblance et il devient, en théorie, possible d'estimer la position de Maximum de Vraisemblance de cette racine le long de l'arbre (Yang and Roberts, 1995). Étant donné qu'il n'y a que quelques études ayant tenté d'analyser cette question, il n'est pas possible actuellement d'avoir une bonne compréhension sur la capacité qu'ont les modèles non-stationnaires à correctement raciner les arbres phylogénétiques. Ainsi, Huelsenbeck et al. (2002) ont montré qu'il fallait de grandes différences compositionnelles pour que le signal de

positionnement de la racine soit détectable. En revanche, Yap and Speed (2005) ont ré-introduit l'intérêt d'utilisation de tels modèles pour raciner des arbres, en se montrant plus optimiste que Huelsenbeck et al. (2002).

Le manuscrit qui suit traite de la phylogénie des Protéobactéries sous l'angle des deux points mentionnés ci-dessus. Nous nous sommes tout d'abord intéressés au potentiel qu'ont les protéines ribosomiques à inférer des arbres phylogénétiques procaryotes à l'échelle du phylum, puis à leur utilisation pour tenter d'inférer la position de la racine de ce phylum bactérien. J'ai participé à cette étude en utilisant le modèle non-homogène et non-stationnaire de Galtier et Gouy (Galtier and Gouy, 1998) implémenté dans nhPhyML (Boussau and Gouy, 2006) pour aborder ces deux problèmes.

Le manuscrit est actuellement soumis.

Remarque : le fichier de matériels supplémentaires associé à cet article est diponible sur demande à cette adresse : `mgroussi@gmail.com`

### 3.3.2 Manuscrit

# Ribosomal proteins as next generation standard for prokaryotic systematics

Hemalatha Golaconda Ramulu[1,¥], Mathieu Groussin[2], Emmanuel Talla[1], Remi Planel[1,‡], Vincent Daubin[2], Céline Brochier-Armanet[1,‡,*]

[1]Aix-Marseille Université; CNRS; UMR 7283; Laboratoire de Chimie Bactérienne, IMM, 31 chemin Joseph Aiguier, F-13402 Marseille, France

[2]Université de Lyon; Université Lyon 1; CNRS; UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 43 boulevard du 11 novembre 1918, F-69622 Villeurbanne, France

Present addresses

[‡]Université de Lyon; Université Lyon 1; CNRS; UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 43 boulevard du 11 novembre 1918, F-69622 Villeurbanne, France

[¥]Aix-Marseille Université; CNRS; UMR 7257, Laboratoire Architecture et Fonction des Macromolécules Biologiques; Campus de Luminy, 163 Avenue de Luminy, F-13288 Marseille CEDEX 09

*Corresponding author (email celine.brochier-armanet@univ-lyon1.fr; tel +33 4 26 23 44 76; fax +33 4 72 43 13 88)

**Keywords.** Non-homogeneous models, Horizontal gene transfer, Root, Endosymbionts, Long branch attraction, SAR11, Zetaproteobacteria, Compositional bias.

**Running title.** R-proteins as a proxy for prokaryotic systematics.

**Abstract.**

The seminal work of Carl Woese and co-workers has contributed to promote the RNA component of the small subunit of the ribosome (SSU rRNA) as a “gold standard” of modern prokaryotic taxonomy and systematics, and an essential tool to explore microbial diversity. Yet, this marker has a limited resolving power, especially at deep phylogenetic depth and can lead to strongly biased trees. The ever-larger number of available complete genomes now calls for a novel standard dataset of robust protein markers that may complement SSU rRNA. In this respect, concatenation of ribosomal proteins (r-proteins) is being growingly used to reconstruct large-scale prokaryotic phylogenies, but the resolving power of these markers for systematic and/or taxonomic purposes has not been specifically tested. Using Proteobacteria as a case study, we show that r-proteins contain a reliable phylogenetic signal at both amino acid and nucleic acid level, which is not blurred by mutational saturation or horizontal gene transfer. The analysis of r-protein supermatrices with accurate evolutionary models allows overcoming most tree reconstruction artefacts resulting from compositional biases and/or fast evolutionary rates. R-proteins harbour a robust phylogenetic signal at a wide range of taxonomic depths, which allows clarifying the relationships among most proteobacterial orders and families, along with the position of several unclassified lineages, suggesting some possible revisions of the current classification. In addition, we investigate the root of the Proteobacteria by considering the time-variation of nucleic acid composition and the signal carried by horizontal gene transfers, two approaches that do not require the use of an outgroup and limit tree reconstruction artefacts. Altogether, our analyses support r-proteins as the next generation standard for prokaryotic taxonomy and systematics.

## 1. Introduction.

Reconstructing the evolutionary relationships among prokaryotic organisms is a major challenge in biology and key to many research fields, including evolution, ecology, and medicine (Gribaldo and Brochier, 2009). Since the seminal work of Carl Woese and colleagues at the end of the 70's (Woese and Fox, 1977), prokaryotic systematics and the exploration of microbial diversity have been relying mainly on phylogenetic analysis of the RNA component of the small subunit of the ribosome (SSU rRNA) (Lopez-Garcia and Moreira, 2008). However, single gene markers (including SSU rRNA) are not able to resolve all phylogenetic relationships with confidence, especially the most ancient ones (Gribaldo and Philippe, 2002). Taking the opportunity of the recent burst of complete genome sequencing projects, new phylogenetic approaches based on gene content, gene order, DNA-string comparison, shared rare genomic events, etc. have been developed (see (Delsuc et al., 2005) and references therein). Among them, special emphasis has been put on the simultaneous analysis of numerous protein coding genes through supertrees or supermatrices, which provide systematically better resolved trees than those based on single markers (including SSU rRNA) (Abby et al., 2012). In the case of prokaryotes, implementing such approaches may be complicated by horizontal gene transfer (HGT) which make the evolutionary history of genes different from that of organisms (Gribaldo and Brochier, 2009; Philippe and Douady, 2003). However, even if HGT is a very important evolutionary process, large scale phylogenetic analyses indicate that a congruent phylogenetic signal reflecting the phylogeny of organisms can be extracted from protein markers (Brochier et al., 2002; Lerat et al., 2003; Matte-Tailliez et al., 2002; Puigbo et al., 2010).

In the post genomic era, the first step of most of studies aiming at investigating the phylogeny of a particular taxonomic group consists in identifying the best-suited set of genes (or proteins) to address the question. Among them, the analysis of the core genome (i.e. orthologous genes present in a single copy per genome) is becoming growingly popular (Kelly et al., 2010; Lang et al., 2013; Touchon et al., 2009; Williams et al., 2010). This is because orthologous are numerous, easy to identify in complete genomes and supposed to

be less affected by HGT, gene duplications and losses. Such approaches have led to significant improvements on our knowledge of the evolutionary history of prokaryotes, in particular by allowing resolving difficult evolutionary issues such as the origin of enterobacteriales obligate endosymbionts of insects (Husnik et al., 2011). Accordingly, we are witnessing a bloom of methodological developments (see (Kuzniar et al., 2008) and references therein) and databases allowing the selection of core genes at different taxonomic levels (DeLuca et al., 2012; Marthey et al., 2008; Ranwez et al., 2007; Wang and Wu, 2013). However, the application of such strategies to systematics or taxonomic purposes deserves careful consideration. Indeed, the identification of core genes strongly depends on the taxonomic sampling of the lineage under study. For instance, the core genome of *Escherichia coli* varies according to the number and the type of strains considered (Touchon et al., 2009). Moreover, core genomes are not comparable from one lineage to another. For example, the core genome of *Escherichia* (Touchon et al., 2009) is different in size and gene content from that of *Mycobacterium* (Tettelin et al., 2005). Finally, the methods (e.g. sequence similarity, single gene phylogeny, etc.) and parameters used to identify orthologous genes/proteins can also strongly influence the delineation of core genomes (Kuzniar et al., 2008). Altogether, differences in gene sampling make the comparison among different studies and the discrepancies observed from one study to the other difficult to interpret from a systematic and/or taxonomic point of view. In addition, the identification of the core genome can be time-consuming, and because it depends on the taxonomic sampling of the lineages under study, it has to be recomputed prior to each analysis. Finally, the combination of core genes can lead to very large supermatrices of characters that can impose a very heavy burden in calculation time, precluding their use to address routinely taxonomic and systematic issues.

There is therefore an urgent need to define a stable and standardized set of molecular markers that overcomes these major issues. These should (i) be largely conserved across prokaryotic lineages, (ii) be easily identifiable in complete genome sequences, (iii) be rarely transferred, and (iv) harbour a robust and reliable phylogenetic signal at various

taxonomic levels, from ancient to more recent relationships. Among protein markers, those involved in translation, and in particular ribosomal proteins (r-proteins), fulfil most of these criteria. Indeed, while a few cases of HGTs have been reported (Brochier et al., 2000; Chen et al., 2009), r-proteins carry a robust phylogenetic signal that can be used to reconstruct ancient phylogenies in the three domains of life (Brochier et al., 2002; Brown et al., 2001; Ciccarelli et al., 2006; Matte-Tailliez et al., 2002; Swithers et al., 2009). However, the resolving power of these markers for systematic and/or taxonomic purposes has not been specifically tested. In this study, we evaluate the value of r-proteins as a proxy for prokaryotic taxonomy and systematics by using Proteobacteria as a case study.

Proteobacteria (from the Greek god “Proteus”, who was capable of assuming many different shapes (Stackebrandt et al., 1988)) represent the largest and phenotypically most diverse bacterial lineage. It encompasses the majority of Gram-negative bacteria, shows a wide diversity of metabolisms and morphologies, and includes a large number of human, animal, and plant symbionts/pathogens of ecological, medical, industrial, and agricultural interest (Kersters et al., 2006). Proteobacteria are also highly relevant from an evolutionary point of view, as the endosymbiosis of the alphaproteobacterial ancestor of mitochondria represents a key step in eukaryogenesis (Embley et al., 2003; Lang et al., 1999). The importance of this phylum is exemplified by the huge number of complete genome sequences in public databases: they represent 40% of bacterial complete genome sequences available at the NCBI in August 2013 (http://www.ncbi.nlm.nih.gov/). Based on SSU rRNA and protein analyses, Proteobacteria have been divided into five classes, which were arbitrarily designated as Alpha, Beta, Gamma, Delta and Epsilon (see (Kersters et al., 2006). Recently, a new candidate division, the Zetaproteobacteria, has been proposed for *Mariprofundus ferrooxydans* PV-1 and JV-1 strains, the first cultivated neutrophilic Fe-oxidizing bacteria, and their uncultivated relatives (Emerson et al., 2007). While proteobacterial classes are well defined, the location of the root of Proteobacteria and the relationships among the orders and families composing each class are not fully understood. In particular, the phylogenetic position of a number of newly described lineages and of fast

evolving species has remained elusive and/or hotly debated. Due to their diversity and abundance, Proteobacteria represent an interesting study case to assess the value of r-proteins as a potential proxy for prokaryotic taxonomy and systematics.

## 2. Material and Methods.

### 2.1 Data set construction

A subset of 472 proteomes representative of the proteobacterial diversity was downloaded at the NCBI (http://www.ncbi.nlm.nih.gov/) (Supplementary Table S1). The sequences were gathered in a local database. The sequences from the recently published genome of Magnetospira sp. QH-2 (Ji et al., 2013) and from the magneto-ovoid strain MO-1 (a second representative of Magnetococcales, Alphaproteobacteria) ongoing project, were kindly provided by Dr Long-Fey Wu (personal communication) and included in the database. The database was screened with BlastP (Altschul et al., 1997) to identify the homologues of the 55 proteobacterial r-proteins (33 LSU and 22 SSU r-proteins) using *Escherichia coli* sequences as seeds. The absence of any r-protein in a given proteome was verified by screening the nucleic acid sequence of the corresponding genome with tBlastn. Accession numbers of retrieved r-proteins sequences are given in the Supplementary Table S1. The nucleotide sequences corresponding to these r-proteins were retrieved from the NCBI.
The 55 resulting datasets were aligned using MUSCLE 3.6 (Edgar, 2004). The use of other programs (i.e. ClustalW (Larkin et al., 2007) and MAFFT (Katoh and Toh, 2008)) provided very similar multiple alignments (not shown). The resulting alignments were used as template to align the corresponding nucleic acid sequences. At this step, the r-protein S1 was discarded due to the presence of numerous repeats preventing the construction of an accurate alignment. The 54 amino acid and nucleic acid alignments were visually inspected and adjusted when necessary using ED (Philippe, 1993). Protein and nucleic acid alignments were trimmed using the NET application from the MUST package.

164 2.2 SUPERMATRIX CONSTRUCTION

165 The trimmed alignments of individual r-proteins were combined to build
166 supermatrices. When a species harboured several copies of a given r-protein, the less
167 divergent homologue was retained. We constructed 100 supermatrices using the 33 LSU r-
168 proteins and 100 supermatrices using the 21 SSU r-proteins by gathering alignments
169 containing an increasing number of unrepresented species (from 0 up to 99). Examination of
170 the resulting supermatrices suggested that for LSU and SSU r-proteins a maximum of ten
171 missing species represented a good compromise between the amount of missing data and
172 the length of the resulting alignments (upper graph, Supplementary Fig. S1). They
173 corresponded to the concatenation of 28 LSU and 20 SSU r-proteins. As expected, the
174 Maximum Likelihood (ML) trees inferred with these two supermatrices showed similar
175 topologies (not shown), confirming that LSU and SSU r-proteins carried a consistent
176 phylogenetic signal. The two supermatrices were therefore combined into a single alignment
177 (FAA-474) representing 5,228 amino acid positions.

178 We constructed a second set of supermatrices using a more restricted taxonomic
179 sampling (137 organisms, Supplementary Table S1). The supermatrices were built by
180 allowing from 0 up to ten missing species per r-protein family. The examination of the
181 resulting supermatrices showed that a maximum of three missing species represented the
182 best compromise between the amount of missing data and the length of the resulting
183 alignments (lower graph, Supplementary Fig. S1). They correspond to the concatenation of
184 27 LSU r-proteins and 19 SSU r-proteins. The ML trees inferred with these two
185 supermatrices were consistent and in agreement with those inferred with the 474 species
186 (not shown). This confirmed that LSU and SSU r-proteins carried a consistent phylogenetic
187 signal and indicated that the reduction of the taxonomic sampling did not bias the
188 phylogenetic signal. The two supermatrices were combined into a single alignment (FAA-
189 137) containing 5,124 amino acid positions. The nucleic acid version of this supermatrix will
190 be referred to as FNT-137. All the datasets are available upon request to CB-A.
191

192 2.3 PHYLOGENETIC ANALYSES

193 ML phylogenies of individual r-proteins were inferred using TreeFinder v2011 (Jobb et
194 al., 2004) with the Le and Gascuel (LG) model (Le and Gascuel, 2008). In order to take into
195 account the heterogeneity of evolutionary rates across sites, we used a gamma distribution
196 with four discrete classes of sites ($\Gamma_4$) and an estimated alpha parameter. The branch
197 robustness of the ML trees was estimated with the non-parametric bootstrap procedure
198 implemented in TreeFinder (100 replicates of the original dataset).

199 ML trees of the supermatrices were inferred with PhyML (Guindon et al., 2010). The
200 best fitted evolutionary models were selected with ProtTest v2.4 (Abascal et al., 2005) for the
201 amino acid supermatrices (FAA-474 and FAA-137) and with TreeFinder v.2011 (AICc
202 criterion) (Jobb et al., 2004) for the nucleic acid supermatrix (FNT-137). The robustness of
203 the FAA-137 and FNT-137 ML trees was estimated by the non-parametric bootstrap
204 procedure implemented in PhyML (100 replicates of the original dataset), whereas the SH-
205 like support was used for FAA-474 ML trees.

206 Additional ML trees of the FNT-137 supermatrix were inferred with the Galtier and
207 Gouy (GG) (Galtier and Gouy, 1998) non-homogeneous model that was recently
208 implemented in nhPhyML (Boussau and Gouy, 2006) in combination with a gamma
209 distribution ($\Gamma_5$). The discrete version of the model was considered in all analyses, with three
210 values of G+C equilibrium content (0.25, 0.5 and 0.75). Thus, for each branch, the best out of
211 the three possible values was determined by maximum likelihood. It is worth noting that
212 nhPhyML requires a rooted tree as a starting point but does not allow topology exploration
213 around the root so that no Nearest Neighbour Interchange (NNI) can be tested between any
214 two lineages present on each side of the root. However, nhPhyML allows topology
215 exploration on each sub-tree surrounding the root. In agreement with the results from this
216 study (see below), we fixed the root position on the branch separating the
217 Epsilonproteobacteria from the other proteobacterial classes to compute ML trees.

218 Bayesian analyses were performed with PhyloBayes 3.3b to investigate the
219 relationships among species within each class (Lartillot et al., 2009). We used the CAT model

in order to take into account across-site heterogeneities in the amino-acid replacement process (Lartillot and Philippe, 2004). For each class (i.e. Alphaproteobacteria, Betaproteobacteria, Deltaproteobacteria, Gammaproteobacteria, and Epsilonproteobacteria), we ran two MCMC chains in parallel with the CAT+$\Gamma_4$ model. The initial 500 trees were discarded as "burn-in". The remaining trees from each chain were used to test for convergence, compute the 50% majority rule consensus tree and the posterior probabilities by sampling one every ten trees. The chains were stopped when the maxdiff and the effective size became lower than 0.3 and greater than 100, respectively. Similar analyses were performed using the Dayhoff4 recoding option (CAT+REC4+$\Gamma_{4)}$. The four Dayhoff's amino acid families corresponded to [(A,G,P,S,T) (D,E,N,Q) (H,K,R) (F,Y,W,I,L,M,V)] plus cysteins treated as missing data (C= ?).

## 2.4 Detection of HGT

We used Prunier (Abby et al., 2010) to search for possible HGT events in individual r-protein trees built with the subset of 137 species. Prunier attempts to find one of the most parsimonious scenarios of HGTs according to a reference phylogeny. We used the ML phylogeny of FAA-137 as a reference tree, because it was previously shown that supermatrix approaches provide good references to detect HGTs in single gene trees (Abby et al., 2012). The branch robustness is taken into account by Prunier in order to minimize the impact of phylogenetic reconstruction errors and the lack of phylogenetic signal. Here, we considered a threshold of 80% bootstrap in individual r-protein ML trees and set the "forward" parameter to 2. Because HGT scenarios depend on the position of the root in the reference tree, we tested the 271 possible roots of the reference phylogeny in order to find the most parsimonious HGT scenario over all protein families. Similar analyses were performed using a more restricted taxonomic sampling (52 species).

## 2.5 Rooting the proteobacterial phylogeny

The non-homogeneous models, such as the GG model implemented in nhPhyML, render the final likelihood of a tree dependant from its root position, contrary to standard homogeneous models. Accordingly, it is possible to use these models to identify the most likely location of the root of an unrooted phylogenetic tree (Yang and Roberts, 1995). We applied this approach to determine the most likely position of the root of Proteobacteria. To do so, we used the ML phylogeny inferred with the FAA-137 supermatrix with the LG+$\Gamma_4$+I model and the first two codon positions of the FNT-137 supermatrix using the GTR+$\Gamma_4$ or the GG+$\Gamma_5$ model. Three species were removed from the analysis because their position in the ML trees was unresolved (*Mariprofundus ferrooxydans* PV-1) or because of huge evolutionary rates ('*Candidatus* (*Ca.*) Hodgkinia cicadicola' and '*Ca*. Carsonella ruddii PV'). Nine putative root positions were tested. The likelihoods of the resulting trees were further compared for statistical significance with the AU test (Shimodaira, 2002) implemented in the CONSEL program (Shimodaira and Hasegawa, 2001). An independent approach based on the pattern of HGTs inferred by Prunier was used to determine the location of the root of Proteobacteria (see below).

### 2.6 Mutational saturation level

The mutational saturation level of FAA-137 was estimated by comparing the evolutionary distance deduced from ML trees inferred with PhyML to the p-distance (i.e. observed divergence) deduced from the multiple alignments between each pair of sequences. A similar analysis was performed for FNT-137 but by considering each of the three codon position separately.

## 3. Results and Discussion.

### 3.1 Taxonomic distribution of r-proteins in Proteobacteria and supermatrix construction

The survey of the proteobacterial proteomes with BlastP highlighted missing r-proteins in many lineages, yet most of these absences corresponded to annotation errors because the corresponding genes can be easily identified in the corresponding genomic sequences with tBlastN (Supplementary Table S1). More precisely, annotation errors were detected in half of the analysed proteomes (240 out of 474), and a few genomes presented more than 10 unannotated r-protein genes. This observation was in agreement with a recent survey of r-proteins in complete prokaryotic genomes (Yutin et al., 2012) and underlined the poor quality of the annotation of some genome sequences. Beside annotation errors, a few r-proteins were truly missing in some proteobacterial lineages (Supplementary Table S1). For instance, L30 and L32 were absent from the genomes of all Epsilonproteobacteria; L34 and L36 were missing in *Mariprofundus ferrooxydans* PV-1 (the only representative of Zetaproteobacteria), whereas L32 was missing in *Magnetococcus marinus* MC-I and its close relative, the magneto-ovoid strain MO-1. Regarding S22, an extremely restricted taxonomic distribution was observed, the protein being present in only 50 closely related species belonging to Enterobacteriales (a gammaproteobacterial order). This indicated that S22, which is associated to stationary phase ribosomes (see (Maki et al., 2000) and references therein), appeared late during the evolution of Proteobacteria. Conversely, a few r-proteins were present in multiple copies in a few genomes. For instance, this is the case for L31 and S21, for which four copies are found in *Escherichia coli* O157:H7 EDL933 and *Burkholderia* (Betaproteobacteria), respectively (Supplementary Table S1). The loss or, alternatively, the presence of multiple copies of some r-proteins in some taxa is puzzling and should be further investigated from a functional point of view. However, to a few exceptions, our results indicated that the majority of the 55 r-proteins were present in a single copy in Proteobacteria and therefore that the set of r-proteins (and thus the ribosome) has not significantly changed during the diversification of this phylum.

The burst of genome sequencing projects allows combining protein markers to investigate the phylogeny of organisms (Delsuc et al., 2005). In the case of r-proteins, their relative short size hinders the use of supertree approaches, especially when large taxonomic

samplings are considered. In contrast, supermatrices have been shown to be particularly well-suited for this type of data (Brochier et al., 2002; Matte-Tailliez et al., 2002). Briefly, this approach consists in combining the alignments of single phylogenetic markers into a single large alignment (called a supermatrix), which is then used for phylogenetic reconstruction (Delsuc et al., 2005). We applied this strategy to build the FAA-474 supermatrix that gathered the 28 LSU and 20 SSU r-proteins presenting a sufficient taxonomic sampling (see methods). The ML tree inferred with this supermatrix recovered the monophyly of most proteobacterial taxa (Supplementary Fig. S2), confirming that r-proteins and SSU rRNA carried an overall consistent phylogenetic signal. Due to biases in the taxonomic distribution of genome projects, some taxa were overrepresented (Supplementary Table S1). To limit taxonomic sampling biases and to reduce computation time, we selected a subset of 137 organisms encompassing most of the taxonomic and genetic diversity of each proteobacterial class to investigate the phylogenetic signal contained in r-proteins and the phylogeny of this bacterial phylum in more detail (species in bold on Supplementary Fig. S2 and Table S1). To do so, we built two supermatrices, FAA-137 and FNT-137, which gathered the amino acid and the nucleic acids alignments of 46 (27 LSU and 19 SSU) r-proteins, respectively.

### 3.2 Protein and nucleic acid sequences of r-proteins contain a reliable phylogenetic signal

The decay of the ancient phylogenetic signal contained in molecular data by successive substitutions occurring at the same position is a frequent problem encountered in phylogeny. This phenomenon is called mutational saturation. Beside the loss of information, mutational saturation may generate tree reconstruction artefacts such as Long Branch Attraction (LBA) which tends to group together sequences associated to long branches (Bergsten, 2005; Felsenstein, 1978; Philippe and Laurent, 1998). This has been extremely well documented in the case of ancient phylogenies (Gribaldo and Philippe, 2002). Because Proteobacteria are an ancient phylum, a certain level of mutational saturation is expected in their r-protein sequences, and thus in the FAA-137 and FNT-137 supermatrices.

330 The level of mutational saturation can be revealed by comparing the p-distances (i.e.
331 the observed substitutions) between each pair of sequences to the corresponding ML-
332 estimated distances. A strong correlation between the two distances was observed in the
333 case of FAA-137 ($R^2$=0.847, Fig. 1a). This suggested that the level of mutational saturation in
334 this supermatrix is moderate and that most of the ancient phylogenetic signal contained in r-
335 proteins was preserved during the diversification of Proteobacteria. Expectedly, the ML and
336 p-distances were strongly correlated among closely related species and/or slowly evolving
337 sequences, whereas the highest discrepancies were observed for pair of sequences
338 involving the two very fast evolving endosymbionts '*Ca.* Hodgkinia cicadicola' and '*Ca*.
339 Carsonella ruddii' (surrounded by a dot line, Fig. 1a), for which more than 3 substitutions per
340 site were inferred with ML, whereas only 0.6 to 0.7 substitutions per site were observed at
341 the sequence level. This indicated that many substitutions occurred in these sequences
342 (large ML-distances), but are hidden due to mutational saturation (moderate p-distances).
343 The analysis of the FNT-137 supermatrix provided a very different picture (Fig. 1b). While a
344 strong correlation was observed between the p- and the ML-distances at the two first codon
345 positions ($R^2$ = 0.9109 and $R^2$ = 0.9431, respectively), a very weak correlation and a great
346 dispersal was observed at the third codon position ($R^2$ = 0.127). This reflects the fast
347 evolutionary rate of the third codon position and its higher saturation with respect to the two
348 other positions. Actually, while a maximum of 4.29 and 3.34 substitutions per site was
349 estimated by ML at the two first codon positions, respectively, up to 10.68 substitutions per
350 site were inferred at the third codon position (Fig. 1b). In addition to higher evolutionary rates,
351 the highest heterogeneity in term of G+C content was observed at the third codon position
352 with respect to the two other positions due to its strong correlation with the genomic G+C
353 content (Fig. 2). These results were expected because the selective pressures are known to
354 be more relaxed at the third codon position due to the redundancy of the genetic code.
355 The combined effect of base composition heterogeneity and fast evolutionary rate
356 may strongly bias tree reconstructions. This was recently illustrated in the case of
357 *Plasmodium* (Davalos and Perkins, 2008) and turtles (Chiari et al., 2012). Proteobacteria are

no exception, as illustrated by the ML phylogeny inferred with the FNT-137 supermatrix using the GTR+$\Gamma_4$ model, which was the best-fitted model proposed by TreeFinder (Supplementary Fig. S3). As all homogeneous and stationary models, the GTR model assumes that the sequences are at equilibrium and thus have the same base composition (see below). Figure 2 shows that this assumption is strongly violated in our data. This may explain the artefactual clustering of unrelated sequences sharing similar base compositions, as illustrated by the grouping of low G+C espilonproteobacteria, low G+C alphaproteobacteria and *Bdellovibrio bacteriovorus* (Deltaproteobacteria) (Supplementary Fig. S3). Expectedly, the monophyly of these classes was recovered when the third codon position was removed from the analysis (Supplementary Fig. S4) or when the protein FAA-137 supermatrix was used (Fig. 3). Importantly, artefactual clustering can also occur at smaller evolutionary scales as exemplified by the grouping of '*Ca*. Liberibacter asiaticus str. psy62' and *Bartonella grahamii* as4aup, two unrelated rhizobiales harbouring moderate G+C contents compared to other representatives of this order of Alphaproteobacteria (Supplementary Fig. S3), which are in fact related to *Sinorhizobium medicae* WSM419 and *Brucella suis* 1330, respectively (Fig. 3 and Supplementary Fig. S4). Expected relationships were also recovered by applying the non-homogeneous Galtier and Gouy (GG) model on the three and on the first two codon positions of the FNT-137 supermatrix (Supplementary Figs. S5 and S6), because this model allows the process of evolution to vary through time and which therefore models variations of base composition among lineages (Galtier and Gouy, 1998). The GG is thus more able to discriminate homoplasies owing to compositional convergence from the true phylogenetic signal than homogeneous models, like GTR. These results strengthened the idea that the use of methods and/or evolutionary models designed to overcome mutational saturation and compositional biases should be systematically considered for the inference of prokaryotic phylogenies, even at small evolutionary scales.

Altogether, our analyses showed that the phylogenetic signal contained in r-proteins has not been completely blurred by mutational saturation and compositional biases. Then, we asked whether this phylogenetic signal reflected the evolutionary history of

Proteobacteria or if it has been obscured by HGTs. To address this question, we investigated the phylogeny of each r-protein in order to quantify the amount of HGT that has affected their evolutionary history. To do so, we used Prunier, a recently developed statistical approach of gene tree reconciliation (Abby et al., 2010). The r-protein S22 was not taken into account due to its very restricted taxonomic distribution (Supplementary Table S1). The analysis of the 53 remaining r-proteins revealed 68 HGT events, representing 1.28 HGTs per protein family in average (Table 1 and Supplementary Table S2). More precisely, 26 r-proteins were devoid of HGT, whereas one, two, three and four HGTs were inferred for 15, six, two and one r-proteins, respectively. In the case of L28, L33, and L36 more than four HGTs (i.e., five, seven, and 19, respectively) were detected. However, for these three proteins Prunier failed to identify a suitable scenario of HGT (Table 1). Ignoring these three potentially artefactual HGT scenarios, 37 HGT events were inferred, representing 0.7 HGT events per gene family in average. This number was roughly twice as low in a similar analysis performed with a sampling of 52 species (not shown). This indicated that HGTs have rarely affected the evolutionary history of r-proteins in Proteobacteria and that most of the topological inconsistences observed in phylogenies of single r-proteins result from a lack of phylogenetic signal and not from HGT. This also confirmed previous studies showing that proteins involved in large complexes are rarely successfully transferred (Cohen et al., 2011; Jain et al., 1999; Leigh et al., 2011) and that r-proteins can be used to investigate the evolutionary history of Proteobacteria.

### 3.3 The root of Proteobacteria

The ML trees inferred with the FAA-137 supermatrix (LG+$\Gamma_4$+I model, Fig. 3) was overall consistent the trees inferred with the FNT-137 supermatrix (GG+$\Gamma_5$ model, Supplementary Figs. S5-S6). As expected, it strongly supported the monophyly of each class (Fig. 3): Epsilonproteobacteria (BV = 100%), Deltaproteobacteria (BV = 93%), Alphaproteobacteria (BV = 93%) and Beta/Gammaproteobacteria (BV = 91%). It also strongly supported the separation of Epsilon- and Deltaproteobacteria from other

proteobacteria (BV = 98%). Regarding Zetaproteobacteria, in the FAA-137 tree, *Mariprofundus ferrooxydans* PV-1 represents a distinct lineage emerging as the sister-group of the clade formed by Alpha, Beta and Gammaproteobacteria. However, the support for this position was weak (BV = 66%, Fig. 3). Actually, according to this phylogeny, we cannot exclude that Zetaproteobacteria could be the sister-lineage of Beta+Gammaproteobacteria or of Alphaproteobacteria. Further experiments using additional markers or the inclusion of new representatives of this class when they become available are needed to precisely determine the position of Zetaproteobacteria with respect to these proteobacteria classes.

Based on protein signatures, it was proposed that the root of Proteobacteria separates Thiobacteria (i.e., Delta and Epsilonproteobacteria) from the three other classes (Gupta, 2000) or that sulphur oxidation performed by Thiobacteria was ancestral (Cavalier-Smith, 2002). Based on molecular phylogenies, either Deltaproteobacteria (Ciccarelli et al., 2006) or Epsilonproteobacteria (Gupta, 2000; Yutin et al., 2012) appeared as the first emerging class, albeit most of the time with non-significant supports. However, these works aimed at reconstructing global phylogenies of Bacteria without addressing specifically the question of the root of Proteobacteria. To our knowledge, the precise location of the root of Proteobacteria has not been carefully investigated and remained to be elucidated. The usual approach to root a phylogenetic tree is based on the use of outgroups. However, this increases the risk of LBA because the branch separating the ingroup from the outgroup is usually longer than the internal branches of the ingroup (Philippe and Laurent, 1998). Non-homogeneous and non-stationary models of evolution, as the GG model used previously, represent an alternative way to root phylogenies without the use of outgroups (Yang and Roberts, 1995). In fact, homogeneous and stationary models, such as GTR or LG assume that the overall sequence composition does not change through time and that the process is at equilibrium from the root to the leaves. These models are reversible, in the sense that there is no direction of evolution along the inferred trees (Felsenstein, 2004), such that the root can be placed wherever on the tree without influencing the likelihood (Yang, 2006). In contrast, non-homogeneous and non-stationary models do not assume the reversibility

hypothesis and assign specific base frequencies to the root so that its position influences the likelihood of the tree (Boussau and Gouy, 2006; Yang and Roberts, 1995). Because of the features of non-homogeneous and non-stationary models mentioned above, it is possible to use them in order to determine the ML position of the root of a tree for which the topology is known. Here, we used this approach to address the question of the root of Proteobacteria. To do so, we considered the three ML topologies inferred with the FAA-137 supermatrix and the LG+$\Gamma_4$+I model (Fig. 3), and with the first two codon positions of the FNT-137 supermatrix using the GTR+$\Gamma_4$ and the GG+$\Gamma_5$ models (Supplementary Figs. S4 and S6). For each topology, nine root positions were tested, corresponding to all possible placements of the root on the internal branches connecting the proteobacterial classes (Table 2). The three topologies provided very similar results, regarding the rank of the nine roots and the conclusions of the AU test. More precisely, the best likelihood was associated to a rooting on the branch separating Epsilonproteobacteria from all other proteobacterial classes. It is worth noting that while four alternative roots were rejected by AU tests (their AU values were below 0.05), a rooting on the branch leading either to Deltaproteobacteria, to Delta+Epsilonproteobacteria, to *Acidithiobacillus*, or to Alpha+Delta+Epsilonproteobacteria, even if less likely, was not significantly statistically rejected (AU values above 0.05, Table 2).

Another approach to root trees without using outgroup is based on the phylogenetic signal carried by HGT. Indeed, it has been shown recently that the pattern of HGT can be exploited to discriminate among putative root positions in species trees (Abby et al., 2010, 2012). We compared the number of HGTs inferred in r-protein families for all of the 271 possible locations of the root in the FAA-137 tree (Supplementary Table S2). Interestingly, the positions of the root that minimized the number of HGT placed Epsilonproteobacteria (or an espilonproteobacterial lineage) as the first diverging lineage within Proteobacteria (Supplementary Table S2). Other rootings implies much more HGT events (Supplementary Fig. S7). Similar results were obtained with a more restricted taxonomic sampling (52 species) of proteobacteria (not shown).

Altogether, the use of non-homogeneous and non-reversible models, and patterns of

HGT favours a rooting of Proteobacteria at the base of Epsilonproteobacteria.

### 3.4 The phylogenetic position of fast evolving proteobacterial lineages

Proteobacteria contain a number of lineages whose phylogenetic position is difficult to determine (Moran et al., 2008). This is, for instance, the case of obligate endosymbionts of insects, such as the gammaproteobacterium '*Ca.* Carsonella ruddii PV' and the alphaproteobacterium '*Ca.* Hodgkinia cicadicola'. '*Ca.* Hodgkinia cicadicola' is an obligate endosymbiont of the cicada *Diceroprocta semicincta* which harbours one of the smallest genome known to date (144 Kb), and is known to be very fast evolving (McCutcheon et al., 2009). Interestingly, while most bacterial symbionts with highly reduced genomes harbour highly A+T rich genomic sequences, the genome of '*Ca.* H. cicadicola' is G+C rich, suggesting that strong selective pressures counteracted the natural mutational bias toward A+T (Van Leuven and McCutcheon, 2012). McCutcheon and collaborators hypothesized a possible relationship between '*Ca.* H. cicadicola' and Rickettsiales, but phylogenetic analyses of SSU rRNA and protein markers favoured a link with Rhizobiales (McCutcheon et al., 2009). The ML trees of FAA-137 and FNT-137 strongly support the former hypothesis because '*Ca.* H. cicadicola' grouped with Rickettsiales and SAR11 (BV > 85%, Fig. 3, and BV > 95, % Supplementary Figs. S4-S6). However, the long branches harboured by these species suggested the possibility of a LBA. To investigate this hypothesis, we reanalysed the phylogeny of Alphaproteobacteria with the CAT model implemented in PhyloBayes (Lartillot and Philippe, 2004), which is less prone to tree reconstruction artefacts such as the LBA (Lartillot et al., 2007). In contrast with ML trees, the Bayesian tree inferred with the CAT+$\Gamma_4$ model strongly rejected the grouping of '*Ca.* H. cicadicola' with Rickettsiales (Posterior Probability (PP) = 0.97, Fig. 4a), the former being displaced to the apical part of the alphaproteobacterial tree (PP = 0.98, Fig. 4a). However, according to this tree, the precise position of '*Ca.* H. cicadicola' relatively to Rhodobacterales, Rhizobiales and Caulobacterales could not be determined (Fig. 4a). A similar result was obtained when the FAA-137 supermatrix was recoded according to the four Dayhoff's amino acid categories (not shown).

498 This confirmed that the grouping of '*Ca.* H. cicadicola' with the Rickettsiales and SAR11 in
499 the FAA-137 and F137-NT ML trees likely resulted from a LBA.
500 Then, we investigated the phylogenetic position of '*Ca.* C. ruddii', a psyllid
501 endosymbiont, which was described as an intermediate evolutionary state between organism
502 and organelle (Tamames et al., 2007). It harbours a very reduced (160 Kb) and G+C poor
503 (16.6%) genome (Nakabachi et al., 2006). Due to extreme evolutionary rates and
504 compositional biases (both at the nucleic and amino acid levels, see above), the
505 phylogenetic position of this bacterium remains uncertain (Williams et al., 2010). In the ML
506 tree of FAA-137, this gammaproteobacterium robustly emerged within Enterobacteriaceae
507 (BV = 91%, Fig. 3) and more precisely within a large group of obligate endosymbionts of
508 insects including *Wigglesworthia* (a symbiont of tsetse flies), *Buchnera* and '*Ca.* Hamiltonella'
509 (two aphid symbionts), '*Ca.* Baumannia' (a symbiont of sharpshooters), and *Blochmannia* (a
510 symbiont of Ants) (BV = 91%, Fig. 3). This suggested that these obligate endosymbionts of
511 insects derived from a single endosymbiosis event, a hypothesis which contradicts a recent
512 phylogenetic analysis suggesting that at least four lineages of obligate endosymbionts of
513 insects emerged independently from free living species during the diversification of
514 Enterobacteriaceae (Husnik et al., 2011). According to this study: (i) *Sodalis*, *Baumannia*,
515 *Blochmannia* and *Wigglesworthia* could be related to *Pectobacterium* and *Dickeya*; (ii)
516 *Buchnera* to a large group encompassing *Erwinia* and *Pantoea*, its closest relatives, but also
517 *Escherichia*, *Salmonella* and other lineages; (iii) *Hamiltonella* and *Regiella* to *Yersinia* and
518 *Serratia*, and (iv) *Riesia* and *Arsenophonus* to *Xenorhabdus*, *Proteus,* and *Photorhabdus*
519 (Husnik et al., 2011). Beside '*Ca*. C. ruddii', our taxonomic sampling, which was not designed
520 to address specifically the question of the origin of obligate endosymbionts of insects,
521 encompassed representatives of the first three groups. Because of their very long branches,
522 the grouping of these obligate endosymbionts in the FAA-137 and FNT-137 tree was suspect
523 and prompted us to investigate the possibility of a LBA. As in the case of '*Ca.* H. cicadicola',
524 we reanalysed the phylogeny of Gammaproteobacteria with the CAT+$\Gamma_4$. In agreement with
525 ML phylogenies, the Bayesian tree strongly support the monophyly of the enterobacteriales

obligate endosymbionts of insects with a significant statistical support (PP = 0.97), to the notable exception of '*Ca.* Carsonella ruddii PV', which was robustly displaced outside of Enterobacteriales (PP = 0.95) and yet grouped with *Thiomicrospira crunogena* XCL-2 and the two sulphur-oxidizing symbionts, albeit with a non-significant support (PP = 0.51) (Fig. 4b). This indicated that the grouping of '*Ca.* Carsonella ruddii PV' with the enterobacteriales obligate endosymbionts of insects in the FAA-137 and FNT-137 ML trees resulted from a LBA. The recoding of the FAA-137 supermatrix according to the four Dayhoff's amino acid families provided similar results but did not allowed clarifying the phylogenetic position of '*Ca.* Carsonella ruddii PV' (not shown). To further investigate the relationships among enterobacteriales obligate endosymbionts of insects, we removed '*Ca.* Carsonella ruddii PV' from the FAA-137 supermatrix. The Bayesian trees inferred with the CAT+$\Gamma_4$ and CAT+REC4+ $\Gamma_4$ were well resolved and overall consistent (Fig. 5a and 5b, respectively). Interestingly, while '*Ca.* Hamiltonella defensa 5AT' emerged with other enterobacteriales obligate endosymbionts of insects in the former (PP = 1, Fig. 5a), it grouped robustly with *Yersinia* when the amino acids were recoded (PP = 1, Fig. 5b), in agreement with the study of Husnik and colleagues (Husnik et al., 2011). This indicated that the grouping of the '*Ca.* Hamiltonella defensa 5AT' with other enterobacteriales obligate endosymbionts of insects in the F137-AA, F137-NT ML trees and in the F137-AA Bayesian phylogeny inferred without amino acid recoding was likely artefactual. Regarding the other endosymbionts, we could not separate *Baumannia*, *Blochmannia*, and *Wigglesworthia* from *Buchnera* (Fig. 5b). This could mean that their separation was artefactual in the study of Husnik, or more likely that our taxonomic sampling was not sufficient to address this question. Indeed, contrarily to the analysis of Husnik et *al.* our analysis did not aim at dissecting in-depth the relationships among obligate endosymbiontic and free living enterobacteriales, which explained our limited taxonomic sampling for this order. However, even with a very restricted taxonomic sampling (10 enterobacteriales species), we showed that the phylogenetic signal carried by r-proteins was sufficient to overcome (at least partially) the LBA resulting from the very fast evolutionary rates of the alphaproteobacterial and gammaproteobacterial obligate endosymbionts of

554 insects.

555

556 3.5 THE EVOLUTIONARY HISTORY WITHIN PROTEOBACTERIAL CLASSES

557 The ML tree of Proteobacteria inferred with FAA-137 and the Bayesian trees of each
558 proteobacterial classes (inferred without and with amino acid recoding) were overall
559 consistent (Figs. 3, 5, 6 and 7), excepted for the position of the fast evolving obligate
560 symbionts (see above). The monophyly of nearly all orders was recovered except within
561 Gammaproteobacteria (see below). More precisely, the relationships within
562 Epsilonproteobacteria were strongly supported, albeit they were not all in agreement with the
563 current taxonomy. Indeed, *Arcobacter butzleri* (Campylobacteraceae) and *Sulfurimonas*
564 *denitrificans* (Helicobacteraceae) grouped robustly with the unclassified
565 epsilonproteobacterium *Sulfurovorum sp.* (BV=87%, PP = 1.0 and 0.96, Figs. 3 and 6a-b,
566 respectively), and not with other Campylobacteraceae or Helicobacteraceae. This was in
567 agreement with the report of genomic similarities shared between *Arcobacter butzleri* and
568 *Sulfurimonas denitrificans* (Miller et al., 2007), and suggested that *Sulfurovorum sp.*, *A.*
569 *butzleri*, and *S. denitrificans* represent a new family within Epsilonproteobacteria, distinct
570 from Campylobacteraceae, Helicobacteraceae, and Nautiliaceae. The main differences
571 among the three trees concerned the position of this group, which formed the sister-lineage
572 of other Campylobacteraceae and Helicobacteraceae in the Bayesian tree inferred without
573 amino acid recoding (PP = 1.0, Fig. 6a) or was more related to Campylobacteraceae when
574 amino acids were recoded (PP = 0.98, Fig. 6b), whereas it position was unresolved in the
575 FAA-137 ML tree (BV < 85%, Fig. 3).

576 Similarly to Epsilonproteobacteria, a robust phylogeny of Deltaproteobacteria
577 emerged from the ML and Bayesian analyses (Figs. 3 and 6c-d). More precisely, the
578 monophyly of Myxococcales was supported (BV = 96%, PP = 1.0 and PP = 0.62) as well as
579 their grouping with Bdellovibrionales (BV = 83%, PP =1.0 in both Bayesian trees).
580 Furthermore, the monophyly of Geobacteraceae, Desulfobacteraceae, Desulfuromonadales,
581 and Desulfovibrionales was strongly recovered (BV ≥ 97% and PP = 1.0 in both Bayesian

582 trees). The monophyly of Desulfovibrionaceae was significantly supported in the ML tree (BV
583 = 97%, Fig. 3) and in the Bayesian phylogeny inferred without amino acid recoding (PP = 1.0,
584 Fig. 6c), whereas *Desulfovibrio salexigens* grouped with *Desulfomicrobium* and
585 *Desulfohalobium*when amino acid were recoded, albeit with a non-significant support (PP =
586 0.90, Fig. 6d), which could reflect an insufficient phylogenetic signal. A lack of phylogenetic
587 signal could also explain the weak support for Desulfobacteriales, whereas the non-
588 monophyly of Syntrophobacterales represented here by *Syntrophus aciditrophicus SB*
589 (Syntrophaceae) and *Syntrophobacter fumaroxidans MPOB* (Syntrophobacteraceae) was not
590 recovered in the ML tree and in the recoded Bayesian tree, albeit with a non-significant
591 support (Figs. 3 and 6d), and strongly rejected in the Bayesian tree inferred without recoding
592 (PP = 1.0, Fig. 6c).

593 In the case of Alphaproteobacteria, the phylogenetic analysis of r-proteins confirmed
594 the close relationship between *Magnetococcus marinus* MC-I and the strain Magneto-ovoid
595 MO-I strain (BV = 100%, Fig. 3, PP = 1.0 and 0.99, Fig. 7) (Lefevre et al., 2009), as well as
596 the early branching of Magnetococcales with respect to other alphaproteobacterial orders
597 (BV = 93%, Fig. 3) (Spring et al., 1998), strengthening the recent proposal that they
598 represent a proteobacterial lineage of high taxonomic rank (Bazylinski et al., 2013). More
599 generally, the monophyly of most alphaproteobacterial orders was recovered (Figs. 3 and 7).
600 One exception concerned Rhodobacterales due to the robust grouping of Caulobacterales
601 with or within Hyphomonadaceae (BV = 100%, PP = 1.0 and 0.99, Figs. 3, 7a and 7b,
602 respectively). Such a clustering has been observed previously in protein and in SSU rRNA
603 trees (Badger et al., 2005; Lee et al., 2005; Thrash et al., 2011; Williams et al., 2007) and is
604 supported by genomic and biological features. This suggested that the boundaries of
605 Caulobacterales and Rhodobacterales should be revised. An interesting point concerned the
606 phylogenetic position of the SAR11 lineage with respect to Rickettsiales. SAR11 represented
607 here by '*Ca.* Pelagibacter ubique' (Giovannoni et al., 2005) is a major component of ocean
608 surface waters (Morris et al., 2002; Rappe et al., 2002; Steindler et al., 2011). The
609 phylogenetic position of SAR11 remained controversial: some analyses suggested that this

group is related to Rickettsiales (Rappe et al., 2002; Thrash et al., 2011) and thus to mitochondria, whereas others supported a relationship with free-living marine and soil alphaproteobacteria and explained the phylogenetic proximity observed between SAR11 and Rickettsiales as the result of compositional biases (Brindefalk et al., 2011; Rodriguez-Ezpeleta and Embley, 2012; Viklund et al., 2011). In the FAA-137 ML tree, Rickettsiales and SAR11 grouped together (BV = 88%) and represented the second diverging order within Alphaproteobacteria (Fig. 3). Interestingly, the grouping of Rickettsiales and SAR11 was strongly rejected in the Bayesian trees inferred with the CAT+$\Gamma_4$ or with the CAT+REC4+$\Gamma_4$ model, the latter being displaced to the apical part of the trees (PP = 1.0 and PP = 0.96, Fig. 7a and 7b, respectively). This indicated that the grouping of SAR11 with Rickettsiales in ML trees resulted from a LBA due to the fast evolutionary rates of these two lineages. However, based on these analyses, we could not precise the position of SAR11 with respect to Sphingomonadales, Rhodobacterales, Caulobacterales and Rhizobiales. To tackle this issue, a broader taxonomic sampling of this lineage and of Alphaproteobacteria in general would be required. However, even with a restricted taxonomic sampling, the phylogenetic signal carried by r-proteins allowed strengthening the hypothesis that SAR11 and Rickettsiales have different origins.

Concerning Betaproteobacteria, the ML phylogeny inferred with FAA-137 and the two Bayesian trees restricted to Betaproteobacteria revealed very few discrepancies (Fig. 3 and 6e-f). In particular, the monophyly of all orders and families was strongly recovered. We confirmed the emergence of '*Ca*. Accumulibacter phosphatis' within Rhodocyclaceae/Rhodocyclales, in agreement with previous studies (Hesselmann et al., 1999), and suggested that it represented a *bona fide* representative of this taxon. The main difference between the three trees concerned the position of *Thiobacillus denitrificans* (Hydrogenophilaceae/Hydrogenophilales), which grouped with Nitrosomonadales in both FAA-137 ML and non-recoded Bayesian trees (BV = 69% and PP = 0.98, Figs. 3 and 6e), whereas it represented an isolated lineage when amino acids were recoded (Fig. 6f). Finally, while Nitrosomonadales formed the sister-group of Burkholderiales and Rhodocyclales in the

638 ML and recoded Bayesian tree (BV = 89% and PP = 1.0, Fig. 3 and 6f), Neisseriales
639 occupied this position in the Bayesian tree inferred without amino acid recoding (Fig. 6e).
640 In contrast to Betaproteobacteria, the phylogeny of Gammaproteobacteria showed
641 strong discrepancies with the current taxonomy. First of all, in the FAA-137 tree,
642 Acidithiobacillales (represented here by *Acidithiobacillus ferrooxidans* str. ATCC 23270)
643 robustly branched-off at the base of the group composed of Beta- and other
644 Gammaproteobacteria (BV = 91%). This position was in agreement with the recent analysis
645 of 356 protein families from 108 gammaproteobacterial proteomes (Williams et al., 2010).
646 Our analysis strengthened this observation and indicated that Acidithiobacillales are neither
647 Beta- nor Gammaproteobacteria, but form a distinct lineage. This means that the taxonomic
648 affiliation of *Acidithiobacillus* (and thus of Acidithiobacillales) to Gammaproteobacteria based
649 on the phylogenetic analysis of a few SSU rRNA sequences using distance methods (Kelly
650 and Wood, 2000) must be reconsidered, either through the creation of a new class or
651 revision of the boundaries between Beta- and Gammaproteobacteria. Strong discrepancies
652 with the current taxonomy were also observed for three major orders: the Alteromonadales,
653 the Pseudomonadales and the Oceanospirillales. These lineages branched-off in the central
654 part of the gammaproteobacterial tree, i.e., after the divergence of Chromatiales,
655 Methylococcales, Cardiobacterales, Xanthomonadales, Legionellales, and Thiotrichales, but
656 before the diversification of Vibrionales, Pasteurellales, Aeromonadales and
657 Enterobacteriales. A careful examination of the FAA-137 ML and Bayesian trees revealed
658 that the Oceanospirillales families (i.e., Alcanivoracaceae, Hahellaceae, Oceanospirillaceae,
659 and Halomonadaceae) formed four unrelated lineages, with the Alcanivoracaceae family
660 being split into two (Figs. 3 and 5). A similar situation was observed for Alteromonadales that
661 formed four distinct lineages and for Pseudomonadales which were split in three unrelated
662 families. These observations were significantly supported by high BV and PP and are in
663 agreement with the recent study of Williams et *al.* (2010). This situation requires urgently in-
664 depth investigations aiming at revisiting the taxonomy of these families and orders. In
665 contrast, the most apical part of the gammaproteobacterial tree was well resolved and in

agreement with the current taxonomy. The monophyly of Enterobacteriales (including the unclassified '*Ca.* Baumannia cicadellinicola'), of Pasteurellales and of Vibrionales was recovered, as well as the sister relationship between Enterobacteriales and Pasteurellales (all BV > 90%, Fig. 3, and all PP = 1.0, Fig. 5). In contrast, the basal part of the phylogeny of Gammaproteobacteria was moderately resolved. While the monophyly of Xanthomonadales, and Legionellales was recovered and well supported in all trees, the monophyly of Thiotrichales, and Chromatiales was weakly supported in the FAA-137 ML tree (Fig. 3), and not recovered in Bayesian trees (Fig. 5). However, in these trees, the unclassified sulphur oxidizing symbionts robustly clustered within *Thiomicrospira* (Thiotrichales) (Figs.3 and 5), suggesting that they belong to the same lineage. Finally, the relationships among the basal branching orders were not resolved, leaving open the question of the early steps of the diversification of Gammaproteobacteria.

## 4. Conclusions

Using Proteobacteria as a case study, we showed that ribosomal proteins represent a promising proxy for prokaryotic taxonomy and systematics: they are highly conserved among prokaryotes, easily identifiable and have been rarely horizontally transferred. Their combination allows assembling relatively large supermatrices, which contain a moderate level of mutation saturation at the protein level but also at the nucleic level, provided that the third codon position is not taken into account. Importantly, the use of accurate evolutionary models allows overcoming most of the tree reconstruction artefacts linked to fast evolving species and/or compositional biases, which are highly problematic in systematic and taxonomy studies. Finally, we showed that the phylogenetic signal contained in r-proteins is a good proxy of the phylogenetic signal contained in larger sets of conserved genes, while allowing applying ML and Bayesian approaches in acceptable computational time. The phylogenies based on r-proteins allowed us to robustly infer the relationships among orders and families within classes, to assign a number of unclassified proteobacterial lineages to existing taxa and to point out a number of discrepancies with the current proteobacterial

694 taxonomy that deserve consideration. Due to their strong conservation across prokaryotes
695 together with the ever increasing availability of complete genome sequences, we anticipate
696 that r-proteins will represent the next generation standard for prokaryotic systematics.
697
698 **Acknowledgments**
699 C.B.-A. was funded by an ATIP from the CNRS. M.G was the recipient of a PhD grant
700 from the French Ministère de l'Education Nationale. H.G.R. was supported by a post-doctoral
701 fellowship from CNRS. R.P. was supported by a grant from the Agence Nationale de la
702 Recherche (ANR-07-BLAN-02). This project was supported by the ANR-07-BLAN-02
703 (BioSuf) and ANR-10-BINF-01-01 (Ancestrome) grants. We thank the PRABI (Pôle Rhône-
704 Alpes de Bioinformatique) for providing computing facilities. We thank Long-Fey Wu for
705 sharing magneto-ovoid strain MO-I unpublished data. We would like to acknowledge Manolo
706 Gouy, Laura Eme, Céline Petitjean, Ji Boyang, Rym Agrebi, and especially Simonetta
707 Gribaldo for stimulating discussions.
708

**Table 1.** Number of HGTs inferred by Prunier in each r-protein family. We only show results inferred for the eleven root positions (out of 271) that minimize the number of HGTs. They correspond to root number 71 to 77, and 91 to 94 according to Supplementary Table S2. These eleven roots are located within Epsilonproteobacteria or in the branch separating Epsilonproteobacteria from other proteobacterial classes. The * designs the three families suspected to yield artefactual scenarios.

| r-protein | Number of HGT | r-protein | Number of HGT | r-protein | Number of HGT |
|---|---|---|---|---|---|
| L10 | 1 | **L28*** | **7** | S13 | 0 |
| L1 | 2 | L29 | 0 | S14 | 0 |
| L11 | 2 | L30 | 0 | S15 | 0 |
| L13 | 1 | L3 | 1 | S16 | 2 |
| L14 | 0 | L31 | 0 | S17 | 0 |
| L15 | 1 | L32 | 0 | S18 | 1 |
| L16 | 1 | **L33*** | **5** | S19 | 0 |
| L17 | 0 | L34 | 0 | S20 | 1 |
| L18 | 0 | L35 | 3 | S2 | 2 |
| L19 | 1 | **L36*** | **19** | S21 | 4 |
| L20 | 0 | L4 | 1 | S3 | 2 |
| L2 | 1 | L5 | 0 | S4 | 3 |
| L21 | 1 | L6 | 0 | S5 | 0 |
| L22 | 0 | L7 | 1 | S6 | 0 |
| L23 | 1 | L9 | 2 | S7 | 0 |
| L24 | 0 | S10 | 0 | S8 | 1 |
| L25 | 1 | S11 | 0 | S9 | 0 |
| L27 | 0 | S12 | 0 | | |

Total number of HGTs: 68 (corresponding to 1.28 HGT per r-protein family)

Total number of HGTs excluding the three doubtful scenarios: 37 (corresponding to 0.7 HGT per r-protein family)

719 **Table 2.** Results of the AU test for the position of the proteobacterial root. The root is located
720 on the branch connecting the two groups separated by the vertical bar. The nine tested
721 positions tested are ranked according to their likelihood computed by nhPhyML. The
722 topologies used as input are the ML trees based on the FAA-137 with the LG+$\Gamma_4$+I model (1)
723 and on the first two codon positions of the FNT-137 supermatrix inferred either with the
724 GTR+$\Gamma_4$ (2) or the GG+$\Gamma_5$ (3) model. E: Epsilonproteobacteria, D: Deltaproteobacteria, T:
725 *Acidithiobacillus*, B: Betaproteobacteria, A: Alphaproteobacteria, G: Gammaproteobacteria.
726

| Position of the root | Rank | ∆Ln*L* (1) | AU *p-value* | Rank | ∆Ln*L* (2) | AU *p-value* | Rank | ∆Ln*L* (3) | AU *p-value* |
|---|---|---|---|---|---|---|---|---|---|
| E\|T,B,G,A,D | 1 | 0 | 0.937 | 1 | 0 | 0.930 | 1 | 0 | 0.961 |
| D\|E,T,B,G,A | 2 | 32.6 | 0.215 | 2 | 32.0 | 0.182 | 2 | 27.6 | 0.246 |
| T\|E,D,B,G,A | 3 | 36.3 | 0.188 | 3 | 36.2 | 0.188 | 3 | 31.3 | 0.169 |
| E,D\|T,B,G,A | 4 | 37.2 | 0.299 | 4 | 37.1 | 0.119 | 4 | 32.2 | 0.087 |
| T,B,G\|E,D,A | 5 | 43.2 | 0.152 | 5 | 39.2 | 0.204 | 5 | 37.5 | 0.118 |
| B,G\|T,E,D,A | 6 | 49.8 | **0.036*** | 7 | 50.2 | 0.059 | 6 | 44.8 | **0.018*** |
| A\|B,G,T,E,D | 7 | 56.2 | 0.092 | 6 | 46.0 | 0.051 | 7 | 51.2 | **0.043*** |
| B\|A,G,T,E,D | 8 | 80.1 | 0.084 | 8 | 51.4 | **0.044*** | 8 | 75.1 | 0.091 |
| G\|T,Ac,B,E,D | 9 | 83.2 | 0.086 | 9 | 51.9 | **0.037*** | 9 | 78.2 | **0.041*** |

727 ***: *p-value* < 0.05, **: *p-value* < 0.01**
728

729 **Legend of figures.**

730

731 **Fig. 1.** Mutational saturation level of (**a**) FAA-137 and (**b**) FNT-137 at the first (black), second
732 (dark grey) and third (light grey) codon positions.

733

734 **Fig. 2.** Distribution of G+C frequencies at the three codon positions **(a-c)** and at the genome
735 level **(d)**. Correlation between the G+C content at each codon position and the whole
736 genome G+C content **(e-g)**.

737

738 **Fig. 3.** ML phylogeny of a subset of 137 Proteobacteria based on the FAA-137supermatrix
739 (5,124 amino acids positions). The tree was inferred with the LG + $\Gamma_4$ + I model, as
740 suggested by ProtTest v2.4. The scale bar represents the average number of substitutions
741 per site. Coloured circles correspond to bootstrap values ranges (100 replicates of the
742 original dataset). For clarity, supports lower than 85% are not shown. Epsilonproteobacteria
743 are in dark blue, Deltaproteobacteria in green-blue, Alphaproteobacteria in pink,
744 Zetaproteobacteria in purple, Betaproteobacteria in light green and Gammaproteobacteria in
745 light blue. Orders and families are depicted in black and in light grey, respectively, according
746 to current taxonomy. Coloured rectangles correspond to genomic G+C contents.

747

748 **Fig. 4.** Bayesian phylogenies of Alphaproteobacteria (**a**) and Gammaproteobacteria (**b**)
749 inferred using the FAA-137 supermatrix (5,124 amino acids positions) with the CAT+$\Gamma_4$
750 model. The scale bars represent the average number of substitutions per site. The statistical
751 supports correspond to posterior probabilities estimated with PhyloBayes. Orders and
752 families are depicted in black and in light grey, respectively, according to current taxonomy.
753 The trees were rooted according to Fig. 3.

754

755 **Fig. 5.** Bayesian phylogenies of Gammaproteobacteria inferred using the FAA-137
756 supermatrix (5,124 amino acids positions) with the CAT+$\Gamma_4$ model (**a**) and the CAT+REC4+$\Gamma_4$

757 model (**b**). The scale bars represent the average number of substitutions per site. The
758 statistical supports correspond to posterior probabilities estimated with PhyloBayes. Orders
759 and families are depicted in black and in light grey, respectively, according to current
760 taxonomy. The trees were rooted according to Fig. 3.
761
762 **Fig. 6.** Bayesian phylogenies of Epsilonproteobacteria (**a** and **b**) and Deltaproteobacteria (**c**
763 and **d**) and Betaproteobacteria (**e** and **f**) inferred using the FAA-137 supermatrix (5,124
764 amino acids positions) with the CAT+$\Gamma_4$ model (**a**, **c** and **e**) and with the CAT+REC4+$\Gamma_4$
765 model (**b**, **d** and **f**). The scale bars represent the average number of substitutions per site.
766 The statistical supports correspond to posterior probabilities estimated with PhyloBayes.
767 Orders and families are depicted in black and in light grey, respectively, according to current
768 taxonomy. The trees were rooted according to Fig. 3.
769
770 **Fig. 7.** Bayesian phylogenies of Alphaproteobacteria inferred using the FAA-137 supermatrix
771 (5,124 amino acids positions) with the CAT+$\Gamma_4$ model (**a**) and the CAT+REC4+$\Gamma_4$ model (**b**).
772 The scale bars represent the average number of substitutions per site. The statistical
773 supports correspond to posterior probabilities estimated with PhyloBayes. Orders and
774 families are depicted in black and in light grey, respectively, according to current taxonomy.
775 The trees were rooted according to Fig. 3.
776
777 **Supplementary material**
778
779 **Supplementary Table S1.** Table showing the accession numbers of r-proteins identified in
780 each of the 474 proteomes under study. Yellow cells correspond to unannotated proteins
781 detected by tBlastN; pink cells correspond to the genomes containing unannotated proteins.
782
783 **Supplementary Table S2.** Table showing the number of HGT inferred with Prunier according
784 to the 271 possible rootings of the FAA-137 tree (with or without taking into account the three

785 families for which Prunier had to reset the parameter forward=2 to 0). The corresponding
786 topologies are shown on the second sheet.
787
788 **Supplementary Fig. S1.**
789 Graphs showing the length (in amino acid positions) of the supermatrices built with the r-
790 proteins from the 474 proteomes (**a**) and a subsampling of 137 proteomes (**b**) according to
791 the number of missing species allowed in each individual r-protein alignment.
792
793 **Supplementary Fig. S2.**
794 Unrooted ML phylogeny of the FAA-474 supermatrix (474 organisms, 5,228 amino acids)
795 inferred using the homogeneous model LG + $\Gamma_4$. The scale bar represents the average
796 number of substitutions per site. Numbers at nodes correspond to SH-like supports inferred
797 by PhyML. For clarity only values > 0.5 are shown. Epsilonproteobacteria are in dark blue,
798 Deltaproteobacteria in green-blue, Alphaproteobacteria in pink, Zetaproteobacteria in purple,
799 Betaproteobacteria in light green and Gammaproteobacteria in light blue. The 137 species
800 selected for more in-depth analyses are in bold.
801
802 **Supplementary Fig. S3.**
803 Unrooted ML phylogeny of the FNT-137 supermatrix. The tree was inferred with the
804 homogeneous model GTR + $\Gamma_4$, the best fitted evolutionary model according to the propose
805 model tool implemented in TreeFinder. The scale bar represents the average number of
806 substitutions per site. Coloured circles correspond to bootstrap values ranges (100 replicates
807 of the original dataset). For clarity, supports lower than 85% are not shown.
808 Epsilonproteobacteria are in dark blue, Deltaproteobacteria in green-blue,
809 Alphaproteobacteria in pink, Zetaproteobacteria in purple, Betaproteobacteria in light green
810 and Gammaproteobacteria in light blue. Coloured rectangles on the left correspond to
811 genomic G+C contents.
812

**Supplementary Fig. S4.**

Unrooted ML phylogeny based on the two first codon positions of the FNT-137 supermatrix. The tree was inferred with the non-homogeneous model GTR + $\Gamma_4$. The scale bar represents the average number of substitutions per site. Coloured circles correspond to bootstrap value ranges (100 replicates of the original dataset). For clarity, supports lower than 85% are not shown. Epsilonproteobacteria are in dark blue, Deltaproteobacteria in green-blue, Alphaproteobacteria in pink, Zetaproteobacteria in purple, Betaproteobacteria in light green, and Gammaproteobacteria in light blue. Orders and families are depicted in black and in light grey, respectively, according to current taxonomy. Coloured rectangles correspond to genomic G+C contents.

**Supplementary Fig. S5.**

Rooted ML phylogeny of the FNT-137 supermatrix. The tree was inferred with the non-homogeneous model GG + $\Gamma_5$. The scale bar represents the average number of substitutions per site. Coloured circles correspond to bootstrap values ranges (100 replicates of the original dataset). For clarity, supports lower than 85% are not shown. Epsilonproteobacteria are in dark blue, Deltaproteobacteria in green-blue, Alphaproteobacteria in pink, Zetaproteobacteria in purple, Betaproteobacteria in light green and Gammaproteobacteria in light blue. Orders and families are depicted in black and in light grey, respectively, according to current taxonomy. Coloured rectangles correspond to genomic G+C contents.

**Supplementary Fig. S6.**

Rooted ML phylogeny based on the two first codon positions of the FNT-137 supermatrix. The tree was inferred with the non-homogeneous model GG + $\Gamma_5$. The scale bar represents the average number of substitutions per site. Coloured circles correspond to bootstrap values ranges (100 replicates of the original dataset). For clarity, supports lower than 85% are not shown. Epsilonproteobacteria are in dark blue, Deltaproteobacteria in green-blue,

Alphaproteobacteria in pink, Zetaproteobacteria in purple, Betaproteobacteria in light green and Gammaproteobacteria in light blue. Orders and families are depicted in black and in light grey, respectively, according to current taxonomy. Coloured rectangles correspond to genomic G+C contents.

**Supplementary Fig. S7.**

Graphs showing the number of HGT inferred with Prunier according to the 271 possible rootings of the FAA-137 phylogeny when all the r-proteins are taken into account (**a**) and when the three datasets for which Prunier failed to find a realist scenario of HGT are removed (**b**). Roots 71 to 77 and 91 to 94 minimize the number of HGT.

852

853 **References**

854

855 Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein
856 evolution. Bioinformatics 21, 2104-2105.
857 Abby, S.S., Tannier, E., Gouy, M., Daubin, V., 2010. Detecting lateral gene transfers by
858 statistical reconciliation of phylogenetic forests. BMC bioinformatics 11, 324.
859 Abby, S.S., Tannier, E., Gouy, M., Daubin, V., 2012. Lateral gene transfer as a support for the
860 tree of life. Proceedings of the National Academy of Sciences of the United States of America
861 109, 4962-4967.
862 Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.,
863 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search
864 programs. Nucleic acids research 25, 3389-3402.
865 Badger, J.H., Eisen, J.A., Ward, N.L., 2005. Genomic analysis of Hyphomonas neptunium
866 contradicts 16S rRNA gene-based phylogenetic analysis: implications for the taxonomy of the
867 orders 'Rhodobacterales' and Caulobacterales. International journal of systematic and
868 evolutionary microbiology 55, 1021-1026.
869 Bazylinski, D.A., Williams, T.J., Lefevre, C.T., Berg, R.J., Zhang, C.L., Bowser, S.S., Dean,
870 A.J., Beveridge, T.J., 2013. Magnetococcus marinus gen. nov., sp. nov., a marine,
871 magnetotactic bacterium that represents a novel lineage (Magnetococcaceae fam. nov.,
872 Magnetococcales ord. nov.) at the base of the Alphaproteobacteria. International journal of
873 systematic and evolutionary microbiology 63, 801-808.
874 Bergsten, J., 2005. A review of long-branch attraction. Cladistics 21, 163-193.
875 Boussau, B., Gouy, M., 2006. Efficient likelihood computations with nonreversible models of
876 evolution. Syst Biol 55, 756-768.
877 Brindefalk, B., Ettema, T.J., Viklund, J., Thollesson, M., Andersson, S.G., 2011. A
878 phylometagenomic exploration of oceanic alphaproteobacteria reveals mitochondrial
879 relatives unrelated to the SAR11 clade. PloS one 6, e24457.

880 Brochier, C., Bapteste, E., Moreira, D., Philippe, H., 2002. Eubacterial phylogeny based on
881 translational apparatus proteins. Trends in genetics : TIG 18, 1-5.
882 Brochier, C., Philippe, H., Moreira, D., 2000. The evolutionary history of ribosomal protein
883 RpS14: horizontal gene transfer at the heart of the ribosome. Trends in genetics : TIG 16,
884 529-533.
885 Brown, J.R., Douady, C.J., Italia, M.J., Marshall, W.E., Stanhope, M.J., 2001. Universal trees
886 based on large combined protein sequence data sets. Nature genetics 28, 281-285.
887 Cavalier-Smith, T., 2002. The neomuran origin of archaebacteria, the negibacterial root of the
888 universal tree and bacterial megaclassification. International journal of systematic and
889 evolutionary microbiology 52, 7-76.
890 Chen, K., Roberts, E., Luthey-Schulten, Z., 2009. Horizontal gene transfer of zinc and non-
891 zinc forms of bacterial ribosomal protein S4. BMC evolutionary biology 9, 179.
892 Chiari, Y., Cahais, V., Galtier, N., Delsuc, F., 2012. Phylogenomic analyses support the
893 position of turtles as the sister group of birds and crocodiles (Archosauria). BMC biology 10,
894 65.
895 Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., Bork, P., 2006. Toward
896 automatic reconstruction of a highly resolved tree of life. Science 311, 1283-1287.
897 Cohen, O., Gophna, U., Pupko, T., 2011. The complexity hypothesis revisited: connectivity
898 rather than function constitutes a barrier to horizontal gene transfer. Molecular biology and
899 evolution 28, 1481-1489.
900 Davalos, L.M., Perkins, S.L., 2008. Saturation and base composition bias explain
901 phylogenomic conflict in Plasmodium. Genomics 91, 433-442.
902 Delsuc, F., Brinkmann, H., Philippe, H., 2005. Phylogenomics and the reconstruction of the
903 tree of life. Nat Rev Genet 6, 361-375.
904 DeLuca, T.F., Cui, J., Jung, J.Y., St Gabriel, K.C., Wall, D.P., 2012. Roundup 2.0: enabling
905 comparative genomics for over 1800 genomes. Bioinformatics 28, 715-716.
906 Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high
907 throughput. Nucleic acids research 32, 1792-1797.

908 Embley, T.M., van der Giezen, M., Horner, D.S., Dyal, P.L., Bell, S., Foster, P.G., 2003.
909 Hydrogenosomes, mitochondria and early eukaryotic evolution. IUBMB life 55, 387-395.
910 Emerson, D., Rentz, J.A., Lilburn, T.G., Davis, R.E., Aldrich, H., Chan, C., Moyer, C.L., 2007.
911 A novel lineage of proteobacteria involved in formation of marine Fe-oxidizing microbial mat
912 communities. PloS one 2, e667.
913 Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively
914 misleading. Syst Zool 27, 401-410.
915 Felsenstein, J., 2004. Inferring phylogenies. Sunderland, Massachusetts.
916 Galtier, N., Gouy, M., 1998. Inferring pattern and process: maximum-likelihood
917 implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic
918 analysis. Mol. Biol. Evol. 15, 871-879.
919 Giovannoni, S.J., Tripp, H.J., Givan, S., Podar, M., Vergin, K.L., Baptista, D., Bibbs, L., Eads,
920 J., Richardson, T.H., Noordewier, M., Rappe, M.S., Short, J.M., Carrington, J.C., Mathur, E.J.,
921 2005. Genome streamlining in a cosmopolitan oceanic bacterium. Science 309, 1242-1245.
922 Gribaldo, S., Brochier, C., 2009. Phylogeny of prokaryotes: does it exist and why should we
923 care? Research in microbiology 160, 513-521.
924 Gribaldo, S., Philippe, H., 2002. Ancient phylogenetic relationships. Theoretical population
925 biology 61, 391-408.
926 Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New
927 algorithms and methods to estimate maximum-likelihood phylogenies: assessing the
928 performance of PhyML 3.0. Syst Biol 59, 307-321.
929 Gupta, R.S., 2000. The phylogeny of proteobacteria: relationships to other eubacterial phyla
930 and eukaryotes. FEMS microbiology reviews 24, 367-402.
931 Hesselmann, R.P., Werlen, C., Hahn, D., van der Meer, J.R., Zehnder, A.J., 1999.
932 Enrichment, phylogenetic analysis and detection of a bacterium that performs enhanced
933 biological phosphate removal in activated sludge. Systematic and applied microbiology 22,
934 454-465.
935 Husnik, F., Chrudimsky, T., Hypsa, V., 2011. Multiple origins of endosymbiosis within the

936 Enterobacteriaceae (gamma-Proteobacteria): convergence of complex phylogenetic
937 approaches. BMC biology 9, 87.
938 Jain, R., Rivera, M.C., Lake, J.A., 1999. Horizontal gene transfer among genomes: The
939 complexity hypothesis. Proceedings of the National Academy of Sciences of the United
940 States of America 96, 3801-3806.
941 Ji, B., Zhang, S.D., Arnoux, P., Rouy, Z., Alberto, F., Philippe, N., Murat, D., Zhang, W.J.,
942 Rioux, J.B., Ginet, N., Sabaty, M., Mangenot, S., Pradel, N., Tian, J., Yang, J., Zhang, L.,
943 Zhang, W., Pan, H., Henrissat, B., Coutinho, P.M., Li, Y., Xiao, T., Medigue, C., Barbe, V.,
944 Pignol, D., Talla, E., Wu, L.F., 2013. Comparative genomic analysis provides insights into the
945 evolution and niche adaptation of marine Magnetospira sp. QH-2 strain. Environmental
946 microbiology.
947 Jobb, G., von Haeseler, A., Strimmer, K., 2004. TREEFINDER: a powerful graphical analysis
948 environment for molecular phylogenetics. BMC evolutionary biology 4, 18.
949 Katoh, K., Toh, H., 2008. Recent developments in the MAFFT multiple sequence alignment
950 program. Briefings in bioinformatics 9, 286-298.
951 Kelly, D.P., Wood, A.P., 2000. Reclassification of some species of *Thiobacillus* to the newly
952 designated genera *Acidithiobacillus* gen. nov., *Halothiobacillus* gen. nov. and
953 *Thermithiobacillus* gen. nov. Int. J. Syst. Evol. Microbiol. 50, 489-500.
954 Kelly, S., Wickstead, B., Gull, K., 2010. Archaeal phylogenomics provides evidence in
955 support of a methanogenic origin of the Archaea and a thaumarchaeal origin for the
956 eukaryotes. Proceedings. Biological sciences / The Royal Society.
957 Kersters, K., Devos, P., Gillis, M., Swings, J., Vandamme, P., Stackebrandt, E., 2006.
958 Introduction to the Proteobacteria. In: Dworkin, M., Falkow, S., Rosenberg, E., Schleifer,
959 K.H., Stackebrandt, E. (Eds.), In the Prokaryotes: A Handbook on the Biology of Bacteria.
960 Springer, New York, pp. 3-37.
961 Kuzniar, A., van Ham, R.C., Pongor, S., Leunissen, J.A., 2008. The quest for orthologs:
962 finding the corresponding gene across genomes. Trends in genetics : TIG 24, 539-551.
963 Lang, B.F., Gray, M.W., Burger, G., 1999. Mitochondrial genome evolution and the origin of

964 eukaryotes. Annu. Rev. Genet. 33, 351-397.
965 Lang, J.M., Darling, A.E., Eisen, J.A., 2013. Phylogeny of bacterial and archaeal genomes
966 using conserved genes: supertrees and supermatrices. PloS one 8, e62510.
967 Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H.,
968 Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G.,
969 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947-2948.
970 Lartillot, N., Brinkmann, H., Philippe, H., 2007. Suppression of long-branch attraction
971 artefacts in the animal phylogeny using a site-heterogeneous model. BMC evolutionary
972 biology 7 Suppl 1, S4.
973 Lartillot, N., Lepage, T., Blanquart, S., 2009. PhyloBayes 3: a Bayesian software package for
974 phylogenetic reconstruction and molecular dating. Bioinformatics 25, 2286-2288.
975 Lartillot, N., Philippe, H., 2004. A Bayesian mixture model for across-site heterogeneities in
976 the amino-acid replacement process. Molecular biology and evolution 21, 1095-1109.
977 Le, S.Q., Gascuel, O., 2008. An improved general amino acid replacement matrix. Molecular
978 biology and evolution 25, 1307-1320.
979 Lee, K.B., Liu, C.T., Anzai, Y., Kim, H., Aono, T., Oyaizu, H., 2005. The hierarchical system of
980 the 'Alphaproteobacteria': description of Hyphomonadaceae fam. nov., Xanthobacteraceae
981 fam. nov. and Erythrobacteraceae fam. nov. International journal of systematic and
982 evolutionary microbiology 55, 1907-1919.
983 Lefevre, C.T., Bernadac, A., Yu-Zhang, K., Pradel, N., Wu, L.F., 2009. Isolation and
984 characterization of a magnetotactic bacterial culture from the Mediterranean Sea.
985 Environmental microbiology 11, 1646-1657.
986 Leigh, J.W., Schliep, K., Lopez, P., Bapteste, E., 2011. Let them fall where they may:
987 congruence analysis in massive phylogenetically messy data sets. Molecular biology and
988 evolution 28, 2773-2785.
989 Lerat, E., Daubin, V., Moran, N.A., 2003. From gene trees to organismal phylogeny in
990 prokaryotes: the case of the gamma-Proteobacteria. PLoS biology 1, E19.
991 Lopez-Garcia, P., Moreira, D., 2008. Tracking microbial biodiversity through molecular and

992 genomic ecology. Research in microbiology 159, 67-73.
993 Maki, Y., Yoshida, H., Wada, A., 2000. Two proteins, YfiA and YhbH, associated with resting
994 ribosomes in stationary phase Escherichia coli. Genes Cells 5, 965-974.
995 Marthey, S., Aguileta, G., Rodolphe, F., Gendrault, A., Giraud, T., Fournier, E., Lopez-
996 Villavicencio, M., Gautier, A., Lebrun, M.H., Chiapello, H., 2008. FUNYBASE: a FUNgal
997 phYlogenomic dataBASE. BMC bioinformatics 9, 456.
998 Matte-Tailliez, O., Brochier, C., Forterre, P., Philippe, H., 2002. Archaeal phylogeny based on
999 ribosomal proteins. Molecular biology and evolution 19, 631-639.
1000 McCutcheon, J.P., McDonald, B.R., Moran, N.A., 2009. Origin of an alternative genetic code
1001 in the extremely small and GC-rich genome of a bacterial symbiont. PLoS genetics 5,
1002 e1000565.
1003 Miller, W.G., Parker, C.T., Rubenfield, M., Mendz, G.L., Wosten, M.M., Ussery, D.W., Stolz,
1004 J.F., Binnewies, T.T., Hallin, P.F., Wang, G., Malek, J.A., Rogosin, A., Stanker, L.H., Mandrell,
1005 R.E., 2007. The complete genome sequence and analysis of the epsilonproteobacterium
1006 Arcobacter butzleri. PloS one 2, e1358.
1007 Moran, N.A., McCutcheon, J.P., Nakabachi, A., 2008. Genomics and evolution of heritable
1008 bacterial symbionts. Annu Rev Genet 42, 165-190.
1009 Morris, R.M., Rappe, M.S., Connon, S.A., Vergin, K.L., Siebold, W.A., Carlson, C.A.,
1010 Giovannoni, S.J., 2002. SAR11 clade dominates ocean surface bacterioplankton
1011 communities. Nature 420, 806-810.
1012 Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H.E., Moran, N.A., Hattori, M.,
1013 2006. The 160-kilobase genome of the bacterial endosymbiont Carsonella. Science 314,
1014 267.
1015 Philippe, H., 1993. MUST, a computer package of Management Utilities for Sequences and
1016 Trees. Nucleic acids research 21, 5264-5272.
1017 Philippe, H., Douady, C.J., 2003. Horizontal gene transfer and phylogenetics. Current opinion
1018 in microbiology 6, 498-505.
1019 Philippe, H., Laurent, J., 1998. How good are deep phylogenetic trees? Current opinion in

1020 genetics & development 8, 616-623.
1021 Puigbo, P., Wolf, Y.I., Koonin, E.V., 2010. The tree and net components of prokaryote
1022 evolution. Genome biology and evolution 2, 745-756.
1023 Ranwez, V., Delsuc, F., Ranwez, S., Belkhir, K., Tilak, M.K., Douzery, E.J., 2007. OrthoMaM:
1024 a database of orthologous genomic markers for placental mammal phylogenetics. BMC
1025 evolutionary biology 7, 241.
1026 Rappe, M.S., Connon, S.A., Vergin, K.L., Giovannoni, S.J., 2002. Cultivation of the
1027 ubiquitous SAR11 marine bacterioplankton clade. Nature 418, 630-633.
1028 Rodriguez-Ezpeleta, N., Embley, T.M., 2012. The SAR11 group of alpha-proteobacteria is not
1029 related to the origin of mitochondria. PloS one 7, e30520.
1030 Shimodaira, H., 2002. An approximately unbiased test of phylogenetic tree selection. Syst
1031 Biol 51, 492-508.
1032 Shimodaira, H., Hasegawa, M., 2001. CONSEL: for assessing the confidence of
1033 phylogenetic tree selection. Bioinformatics 17, 1246-1247.
1034 Spring, S., Lins, U., Amann, R., Schleifer, K.H., Ferreira, L.C., Esquivel, D.M., Farina, M.,
1035 1998. Phylogenetic affiliation and ultrastructure of uncultured magnetic bacteria with
1036 unusually large magnetosomes. Archives of microbiology 169, 136-147.
1037 Stackebrandt, E., Murray, R.G.E., Trüper, H.G., 1988. *Proteobacteria* classis nov., a name for
1038 the phylogenetic taxon that includes the "purple bacteria and their relatives". Int. J. Syst.
1039 Bacteriol 38, 321-325.
1040 Steindler, L., Schwalbach, M.S., Smith, D.P., Chan, F., Giovannoni, S.J., 2011. Energy
1041 starved Candidatus Pelagibacter ubique substitutes light-mediated ATP production for
1042 endogenous carbon respiration. PloS one 6, e19725.
1043 Swithers, K.S., Gogarten, J.P., Fournier, G.P., 2009. Trees in the web of life. Journal of
1044 biology 8, 54.
1045 Tamames, J., Gil, R., Latorre, A., Pereto, J., Silva, F.J., Moya, A., 2007. The frontier between
1046 cell and organelle: genome analysis of Candidatus Carsonella ruddii. BMC evolutionary
1047 biology 7, 181.

1048 Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V.,
1049 Crabtree, J., Jones, A.L., Durkin, A.S., Deboy, R.T., Davidsen, T.M., Mora, M., Scarselli, M.,
1050 Margarit y Ros, I., Peterson, J.D., Hauser, C.R., Sundaram, J.P., Nelson, W.C., Madupu, R.,
1051 Brinkac, L.M., Dodson, R.J., Rosovitz, M.J., Sullivan, S.A., Daugherty, S.C., Haft, D.H.,
1052 Selengut, J., Gwinn, M.L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins,
1053 K., O'Connor, K.J., Smith, S., Utterback, T.R., White, O., Rubens, C.E., Grandi, G., Madoff,
1054 L.C., Kasper, D.L., Telford, J.L., Wessels, M.R., Rappuoli, R., Fraser, C.M., 2005. Genome
1055 analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the
1056 microbial "pan-genome". Proceedings of the National Academy of Sciences of the United
1057 States of America 102, 13950-13955.
1058 Thrash, J.C., Boyd, A., Huggett, M.J., Grote, J., Carini, P., Yoder, R.J., Robbertse, B.,
1059 Spatafora, J.W., Rappe, M.S., Giovannoni, S.J., 2011. Phylogenomic evidence for a common
1060 ancestor of mitochondria and the SAR11 clade. Scientific reports 1, 13.
1061 Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E.,
1062 Bonacorsi, S., Bouchier, C., Bouvet, O., Calteau, A., Chiapello, H., Clermont, O., Cruveiller,
1063 S., Danchin, A., Diard, M., Dossat, C., Karoui, M.E., Frapy, E., Garry, L., Ghigo, J.M., Gilles,
1064 A.M., Johnson, J., Le Bouguenec, C., Lescat, M., Mangenot, S., Martinez-Jehanne, V., Matic,
1065 I., Nassif, X., Oztas, S., Petit, M.A., Pichon, C., Rouy, Z., Ruf, C.S., Schneider, D., Tourret, J.,
1066 Vacherie, B., Vallenet, D., Medigue, C., Rocha, E.P., Denamur, E., 2009. Organised genome
1067 dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS
1068 genetics 5, e1000344.
1069 Van Leuven, J.T., McCutcheon, J.P., 2012. An AT mutational bias in the tiny GC-rich
1070 endosymbiont genome of Hodgkinia. Genome biology and evolution 4, 24-27.
1071 Viklund, J., Ettema, T.J., Andersson, S.G., 2011. Independent Genome Reduction and
1072 Phylogenetic Reclassification of the Oceanic SAR11 Clade. Molecular biology and evolution.
1073 Wang, Z., Wu, M., 2013. A phylum-level bacterial phylogenetic marker database. Molecular
1074 biology and evolution 30, 1258-1262.
1075 Williams, K.P., Gillespie, J.J., Sobral, B.W., Nordberg, E.K., Snyder, E.E., Shallom, J.M.,

1076 Dickerman, A.W., 2010. Phylogeny of gammaproteobacteria. Journal of bacteriology 192,
1077 2305-2314.
1078 Williams, K.P., Sobral, B.W., Dickerman, A.W., 2007. A robust species tree for the
1079 alphaproteobacteria. Journal of bacteriology 189, 4578-4586.
1080 Woese, C.R., Fox, G.E., 1977. Phylogenetic structure of the prokaryotic domain: the primary
1081 kingdoms. Proc. Natl. Acad. Sci. USA 74, 5088-5090.
1082 Yang, Z., 2006. Computational Molecular Evolution. Oxford University Press, Oxford,
1083 England.
1084 Yang, Z., Roberts, D., 1995. On the use of nucleic acid sequences to infer early branchings
1085 in the tree of life. Molecular biology and evolution 12, 451-458.
1086 Yutin, N., Puigbò, P., Koonin, E.V., Wolf, Y.I., 2012. Phylogenomics of Prokaryotic Ribosomal
1087 Proteins. PloS one 7, e36972.
1088
1089

**Figure 1**

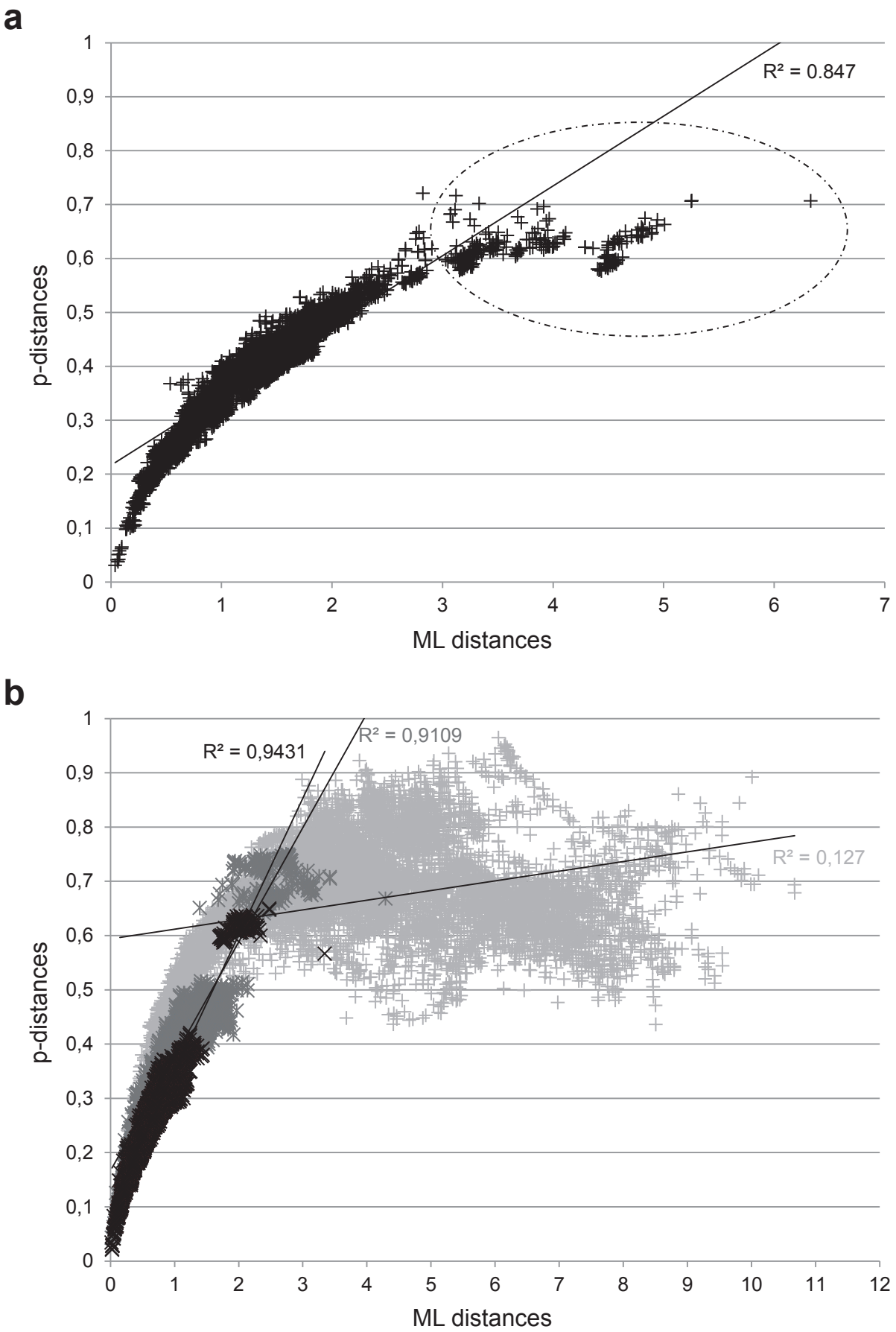

a
R² = 0.847
p-distances
ML distances
b
R² = 0,9431
R² = 0,9109
R² = 0,127
p-distances
ML distances

Figure 2

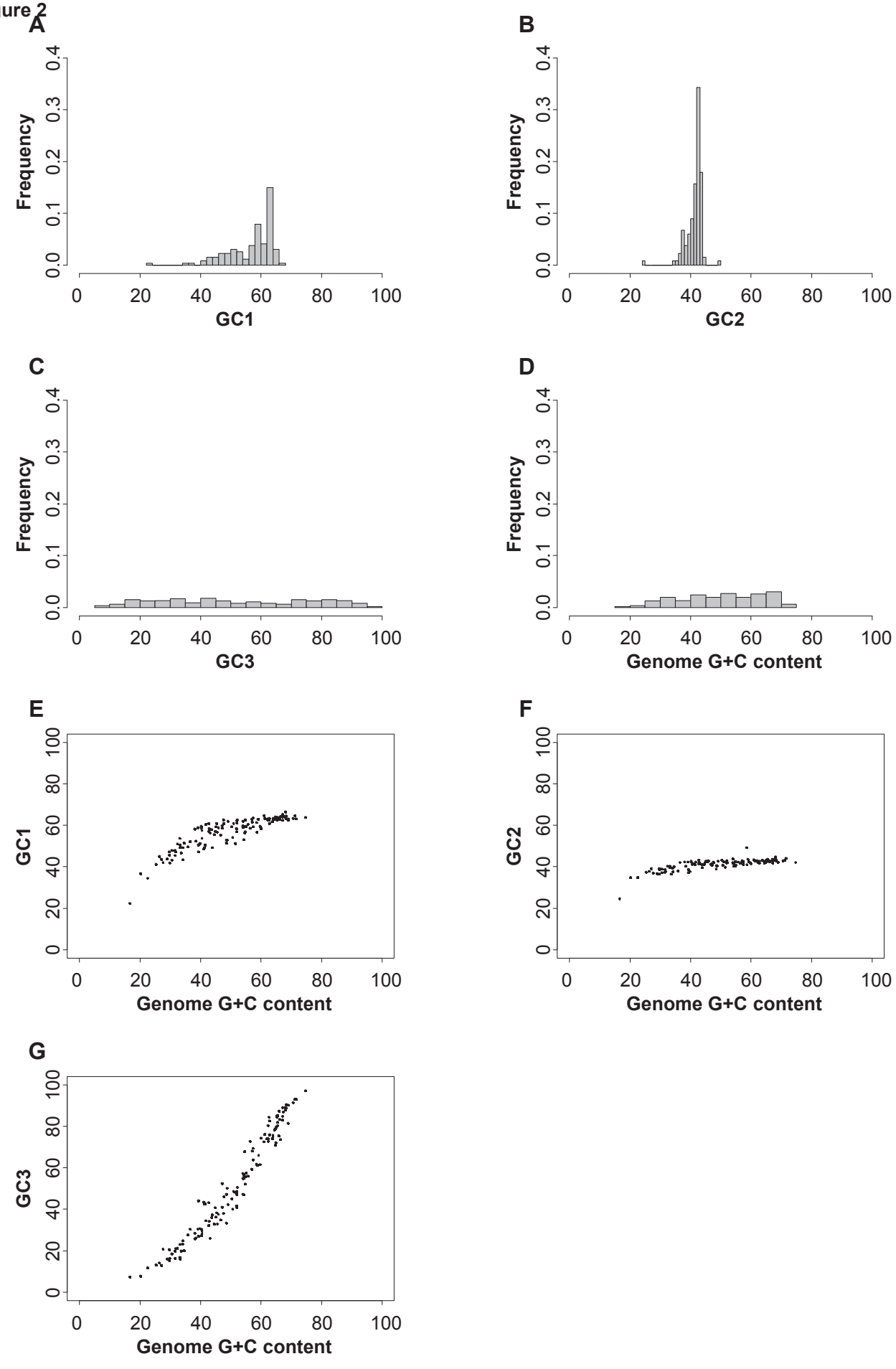

A
Frequency
0.0
0.1
0.2
0.3
0.4
0
20
40
60
80
100
GC1
B
Frequency
GC2
C
Frequency
GC3
D
Frequency
Genome G+C content
E
GC1
Genome G+C content
F
GC2
Genome G+C content
G
GC3
Genome G+C content

**Figure 3**

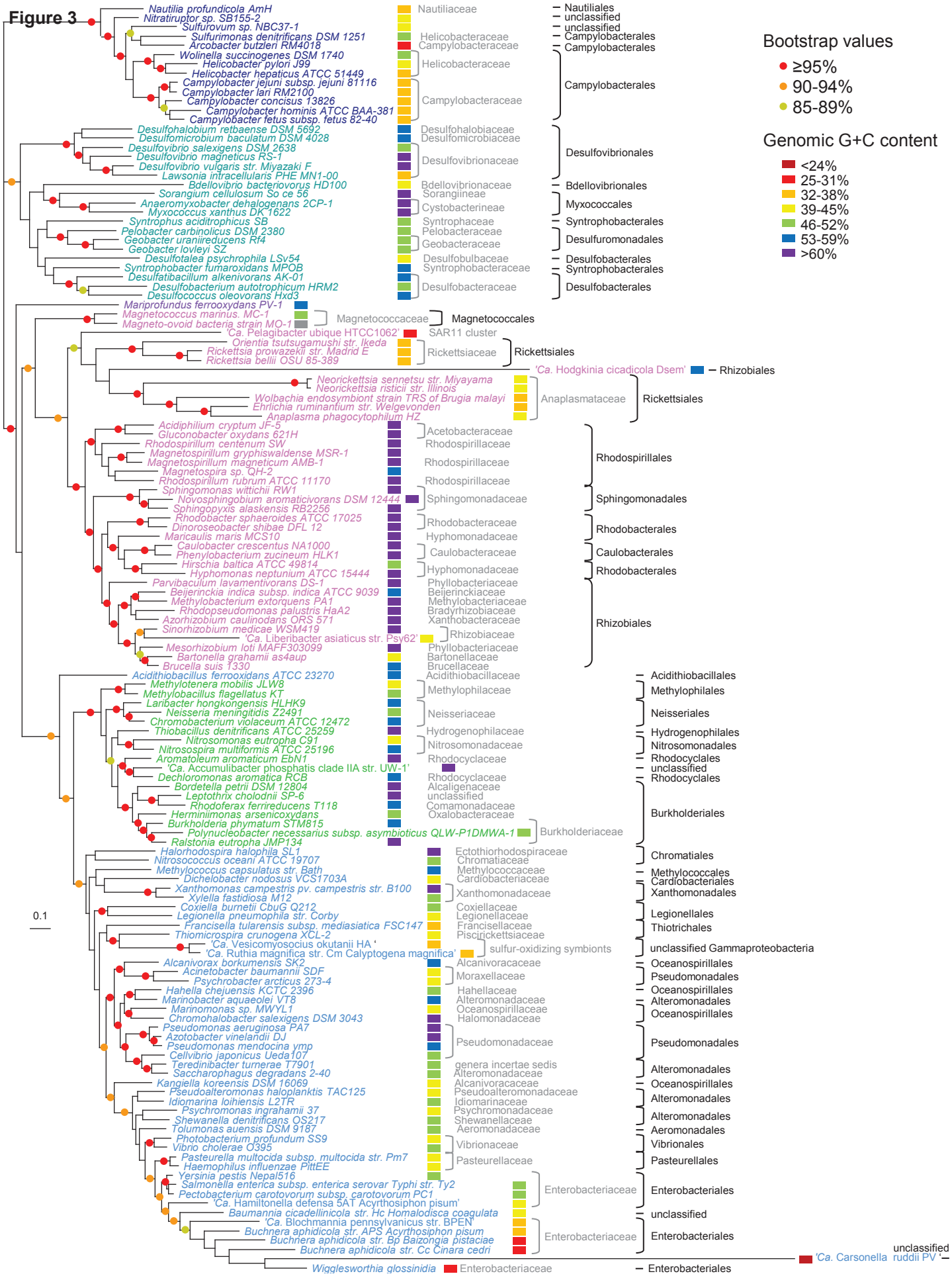

Nautilia profundicola AmH
Nitratiruptor sp. SB155-2
Sulfurovum sp. NBC37-1
Sulfurimonas denitrificans DSM 1251
Arcobacter butzleri RM4018
Wolinella succinogenes DSM 1740
Helicobacter pylori J99
Helicobacter hepaticus ATCC 51449
Campylobacter jejuni subsp. jejuni 81116
Campylobacter lari RM2100
Campylobacter concisus 13826
Campylobacter hominis ATCC BAA-381
Campylobacter fetus subsp. fetus 82-40
Desulfohalobium retbaense DSM 5692
Desulfomicrobium baculatum DSM 4028
Desulfovibrio salexigens DSM 2638
Desulfovibrio magneticus RS-1
Desulfovibrio vulgaris str. Miyazaki F
Lawsonia intracellularis PHE MN1-00
Bdellovibrio bacteriovorus HD100
Sorangium cellulosum So ce 56
Anaeromyxobacter dehalogenans 2CP-1
Myxococcus xanthus DK 1622
Syntrophus aciditrophicus SB
Pelobacter carbinolicus DSM 2380
Geobacter uraniireducens Rf4
Geobacter lovleyi SZ
Desulfotalea psychrophila LSv54
Syntrophobacter fumaroxidans MPOB
Desulfatibacillum alkenivorans AK-01
Desulfobacterium autotrophicum HRM2
Desulfococcus oleovorans Hxd3
Mariprofundus ferrooxydans PV-1
Magnetococcus marinus. MC-1
Magneto-ovoid bacteria strain MO-1
'Ca. Pelagibacter ubique HTCC1062'
Orientia tsutsugamushi str. Ikeda
Rickettsia prowazekii str. Madrid E
Rickettsia bellii OSU 85-389
'Ca. Hodgkinia cicadicola Dsem'
Neorickettsia sennetsu str. Miyayama
Neorickettsia risticii str. Illinois
Wolbachia endosymbiont strain TRS of Brugia malayi
Ehrlichia ruminantium str. Welgevonden
Anaplasma phagocytophilum HZ
Acidiphilium cryptum JF-5
Gluconobacter oxydans 621H
Rhodospirillum centenum SW
Magnetospirillum gryphiswaldense MSR-1
Magnetospirillum magneticum AMB-1
Magnetospira sp. QH-2
Rhodospirillum rubrum ATCC 11170
Sphingomonas wittichii RW1
Novosphingobium aromaticivorans DSM 12444
Sphingopyxis alaskensis RB2256
Rhodobacter sphaeroides ATCC 17025
Dinoroseobacter shibae DFL 12
Maricaulis maris MCS10
Caulobacter crescentus NA1000
Phenylobacterium zucineum HLK1
Hirschia baltica ATCC 49814
Hyphomonas neptunium ATCC 15444
Parvibaculum lavamentivorans DS-1
Beijerinckia indica subsp. indica ATCC 9039
Methylobacterium extorquens PA1
Rhodopseudomonas palustris HaA2
Azorhizobium caulinodans ORS 571
Sinorhizobium medicae WSM419
'Ca. Liberibacter asiaticus str. Psy62'
Mesorhizobium loti MAFF303099
Bartonella grahamii as4aup
Brucella suis 1330
Acidithiobacillus ferrooxidans ATCC 23270
Methylotenera mobilis JLW8
Methylobacillus flagellatus KT
Laribacter hongkongensis HLHK9
Neisseria meningitidis Z2491
Chromobacterium violaceum ATCC 12472
Thiobacillus denitrificans ATCC 25259
Nitrosomonas eutropha C91
Nitrosospira multiformis ATCC 25196
Aromatoleum aromaticum EbN1
'Ca. Accumulibacter phosphatis clade IIA str. UW-1'
Dechloromonas aromatica RCB
Bordetella petrii DSM 12804
Leptothrix cholodnii SP-6
Rhodoferax ferrireducens T118
Herminiimonas arsenicoxydans
Burkholderia phymatum STM815
Polynucleobacter necessarius subsp. asymbioticus QLW-P1DMWA-1
Ralstonia eutropha JMP134
Halorhodospira halophila SL1
Nitrosococcus oceani ATCC 19707
Methylococcus capsulatus str. Bath
Dichelobacter nodosus VCS1703A
Xanthomonas campestris pv. campestris str. B100
Xylella fastidiosa M12
Coxiella burnetii CbuG Q212
Legionella pneumophila str. Corby
Francisella tularensis subsp. mediasiatica FSC147
Thiomicrospira crunogena XCL-2
'Ca. Vesicomyosocius okutanii HA '
'Ca. Ruthia magnifica str. Cm Calyptogena magnifica'
Alcanivorax borkumensis SK2
Acinetobacter baumannii SDF
Psychrobacter arcticus 273-4
Hahella chejuensis KCTC 2396
Marinobacter aquaeolei VT8
Marinomonas sp. MWYL1
Chromohalobacter salexigens DSM 3043
Pseudomonas aeruginosa PA7
Azotobacter vinelandii DJ
Pseudomonas mendocina ymp
Cellvibrio japonicus Ueda107
Teredinibacter turnerae T7901
Saccharophagus degradans 2-40
Kangiella koreensis DSM 16069
Pseudoalteromonas haloplanktis TAC125
Idiomarina loihiensis L2TR
Psychromonas ingrahamii 37
Shewanella denitrificans OS217
Tolumonas auensis DSM 9187
Photobacterium profundum SS9
Vibrio cholerae O395
Pasteurella multocida subsp. multocida str. Pm7
Haemophilus influenzae PittEE
Yersinia pestis Nepal516
Salmonella enterica subsp. enterica serovar Typhi str. Ty2
Pectobacterium carotovorum subsp. carotovorum PC1
'Ca. Hamiltonella defensa 5AT Acyrthosiphon pisum'
Baumannia cicadellinicola str. Hc Homalodisca coagulata
'Ca. Blochmannia pennsylvanicus str. BPEN'
Buchnera aphidicola str. APS Acyrthosiphon pisum
Buchnera aphidicola str. Bp Baizongia pistaciae
Buchnera aphidicola str. Cc Cinara cedri
Wigglesworthia glossinidia
'Ca. Carsonella ruddii PV '
0.1
Nautiliaceae
Helicobacteraceae
Campylobacteraceae
Desulfohalobiaceae
Desulfomicrobiaceae
Desulfovibrionaceae
Bdellovibrionaceae
Sorangiineae
Cystobacterineae
Syntrophaceae
Pelobacteraceae
Geobacteraceae
Desulfobulbaceae
Syntrophobacteraceae
Desulfobacteraceae
Magnetococcaceae
SAR11 cluster
Rickettsiaceae
Anaplasmataceae
Acetobacteraceae
Rhodospirillaceae
Sphingomonadaceae
Rhodobacteraceae
Hyphomonadaceae
Caulobacteraceae
Phyllobacteriaceae
Beijerinckiaceae
Methylobacteriaceae
Bradyrhizobiaceae
Xanthobacteraceae
Rhizobiaceae
Brucellaceae
Acidithiobacillaceae
Methylophilaceae
Neisseriaceae
Hydrogenophilaceae
Nitrosomonadaceae
Rhodocyclaceae
Alcaligenaceae
unclassified
Comamonadaceae
Oxalobacteraceae
Burkholderiaceae
Ectothiorhodospiraceae
Chromatiaceae
Methylococcaceae
Cardiobacteriaceae
Xanthomonadaceae
Coxiellaceae
Legionellaceae
Francisellaceae
Piscirickettsiaceae
sulfur-oxidizing symbionts
Alcanivoracaceae
Moraxellaceae
Hahellaceae
Alteromonadaceae
Oceanospirillaceae
Halomonadaceae
Pseudomonadaceae
genera incertae sedis
Alteromonadaceae
Alcanivoracaceae
Pseudoalteromonadaceae
Idiomarinaceae
Psychromonadaceae
Shewanellaceae
Aeromonadaceae
Vibrionaceae
Pasteurellaceae
Enterobacteriaceae
Enterobacteriaceae
Enterobacteriaceae
Nautiliales
unclassified
unclassified
Campylobacterales
Campylobacterales
Campylobacterales
Desulfovibrionales
Bdellovibrionales
Myxococcales
Syntrophobacterales
Desulfuromonadales
Desulfobacterales
Syntrophobacterales
Desulfobacterales
Magnetococcales
Rickettsiales
Rhizobiales
Rickettsiales
Rhodospirillales
Sphingomonadales
Rhodobacterales
Caulobacterales
Rhodobacterales
Rhizobiales
Acidithiobacillales
Methylophilales
Neisseriales
Hydrogenophilales
Nitrosomonadales
Rhodocyclales
unclassified
Rhodocyclales
Burkholderiales
Chromatiales
Methylococcales
Cardiobacteriales
Xanthomonadales
Legionellales
Thiotrichales
unclassified Gammaproteobacteria
Oceanospirillales
Pseudomonadales
Oceanospirillales
Alteromonadales
Oceanospirillales
Pseudomonadales
Alteromonadales
Oceanospirillales
Alteromonadales
Alteromonadales
Aeromonadales
Vibrionales
Pasteurellales
Enterobacteriales
unclassified
Enterobacteriales
unclassified
Enterobacteriales
Bootstrap values
≥95%
90-94%
85-89%
Genomic G+C content
<24%
25-31%
32-38%
39-45%
46-52%
53-59%
>60%

Figure 4

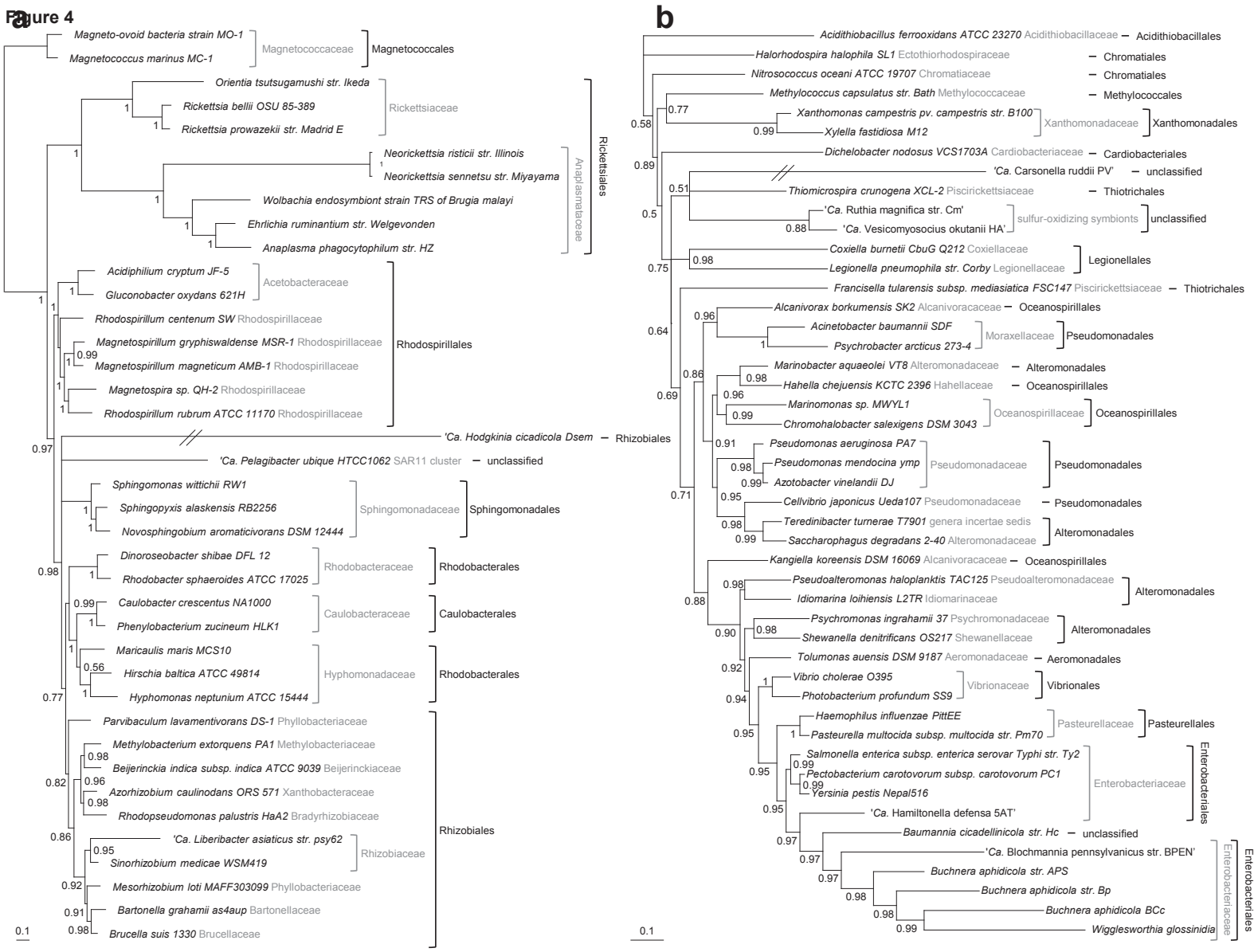

a
Magneto-ovoid bacteria strain MO-1
Magnetococcus marinus MC-1
Magnetococcaceae
Magnetococcales
Orientia tsutsugamushi str. Ikeda
Rickettsia bellii OSU 85-389
Rickettsia prowazekii str. Madrid E
Rickettsiaceae
Neorickettsia risticii str. Illinois
Neorickettsia sennetsu str. Miyayama
Wolbachia endosymbiont strain TRS of Brugia malayi
Ehrlichia ruminantium str. Welgevonden
Anaplasma phagocytophilum str. HZ
Anaplasmataceae
Rickettsiales
Acidiphilium cryptum JF-5
Gluconobacter oxydans 621H
Acetobacteraceae
Rhodospirillum centenum SW Rhodospirillaceae
Magnetospirillum gryphiswaldense MSR-1 Rhodospirillaceae
Magnetospirillum magneticum AMB-1 Rhodospirillaceae
Magnetospira sp. QH-2 Rhodospirillaceae
Rhodospirillum rubrum ATCC 11170 Rhodospirillaceae
Rhodospirillales
'Ca. Hodgkinia cicadicola Dsem
Rhizobiales
'Ca. Pelagibacter ubique HTCC1062 SAR11 cluster
unclassified
Sphingomonas wittichii RW1
Sphingopyxis alaskensis RB2256
Novosphingobium aromaticivorans DSM 12444
Sphingomonadaceae
Sphingomonadales
Dinoroseobacter shibae DFL 12
Rhodobacter sphaeroides ATCC 17025
Rhodobacteraceae
Rhodobacterales
Caulobacter crescentus NA1000
Phenylobacterium zucineum HLK1
Caulobacteraceae
Caulobacterales
Maricaulis maris MCS10
Hirschia baltica ATCC 49814
Hyphomonas neptunium ATCC 15444
Hyphomonadaceae
Rhodobacterales
Parvibaculum lavamentivorans DS-1 Phyllobacteriaceae
Methylobacterium extorquens PA1 Methylobacteriaceae
Beijerinckia indica subsp. indica ATCC 9039 Beijerinckiaceae
Azorhizobium caulinodans ORS 571 Xanthobacteraceae
Rhodopseudomonas palustris HaA2 Bradyrhizobiaceae
'Ca. Liberibacter asiaticus str. psy62
Sinorhizobium medicae WSM419
Rhizobiaceae
Mesorhizobium loti MAFF303099 Phyllobacteriaceae
Bartonella grahamii as4aup Bartonellaceae
Brucella suis 1330 Brucellaceae
Rhizobiales
0.1
b
Acidithiobacillus ferrooxidans ATCC 23270 Acidithiobacillaceae
Acidithiobacillales
Halorhodospira halophila SL1 Ectothiorhodospiraceae
Chromatiales
Nitrosococcus oceani ATCC 19707 Chromatiaceae
Chromatiales
Methylococcus capsulatus str. Bath Methylococcaceae
Methylococcales
Xanthomonas campestris pv. campestris str. B100
Xylella fastidiosa M12
Xanthomonadaceae
Xanthomonadales
Dichelobacter nodosus VCS1703A Cardiobacteriaceae
Cardiobacteriales
'Ca. Carsonella ruddii PV'
unclassified
Thiomicrospira crunogena XCL-2 Piscirickettsiaceae
Thiotrichales
'Ca. Ruthia magnifica str. Cm'
'Ca. Vesicomyosocius okutanii HA'
sulfur-oxidizing symbionts
unclassified
Coxiella burnetii CbuG Q212 Coxiellaceae
Legionella pneumophila str. Corby Legionellaceae
Legionellales
Francisella tularensis subsp. mediasiatica FSC147 Piscirickettsiaceae
Thiotrichales
Alcanivorax borkumensis SK2 Alcanivoracaceae
Oceanospirillales
Acinetobacter baumannii SDF
Psychrobacter arcticus 273-4
Moraxellaceae
Pseudomonadales
Marinobacter aquaeolei VT8 Alteromonadaceae
Alteromonadales
Hahella chejuensis KCTC 2396 Hahellaceae
Oceanospirillales
Marinomonas sp. MWYL1
Chromohalobacter salexigens DSM 3043
Oceanospirillaceae
Oceanospirillales
Pseudomonas aeruginosa PA7
Pseudomonas mendocina ymp
Azotobacter vinelandii DJ
Pseudomonadaceae
Pseudomonadales
Cellvibrio japonicus Ueda107 Pseudomonadaceae
Pseudomonadales
Teredinibacter turnerae T7901 genera incertae sedis
Saccharophagus degradans 2-40 Alteromonadaceae
Alteromonadales
Kangiella koreensis DSM 16069 Alcanivoracaceae
Oceanospirillales
Pseudoalteromonas haloplanktis TAC125 Pseudoalteromonadaceae
Idiomarina loihiensis L2TR Idiomarinaceae
Alteromonadales
Psychromonas ingrahamii 37 Psychromonadaceae
Shewanella denitrificans OS217 Shewanellaceae
Alteromonadales
Tolumonas auensis DSM 9187 Aeromonadaceae
Aeromonadales
Vibrio cholerae O395
Photobacterium profundum SS9
Vibrionaceae
Vibrionales
Haemophilus influenzae PittEE
Pasteurella multocida subsp. multocida str. Pm70
Pasteurellaceae
Pasteurellales
Salmonella enterica subsp. enterica serovar Typhi str. Ty2
Pectobacterium carotovorum subsp. carotovorum PC1
Yersinia pestis Nepal516
Enterobacteriaceae
Enterobacteriales
'Ca. Hamiltonella defensa 5AT'
Baumannia cicadellinicola str. Hc
unclassified
'Ca. Blochmannia pennsylvanicus str. BPEN'
Buchnera aphidicola str. APS
Buchnera aphidicola str. Bp
Buchnera aphidicola BCc
Wigglesworthia glossinidia
Enterobacteriaceae
Enterobacteriales
0.1

Figure 5

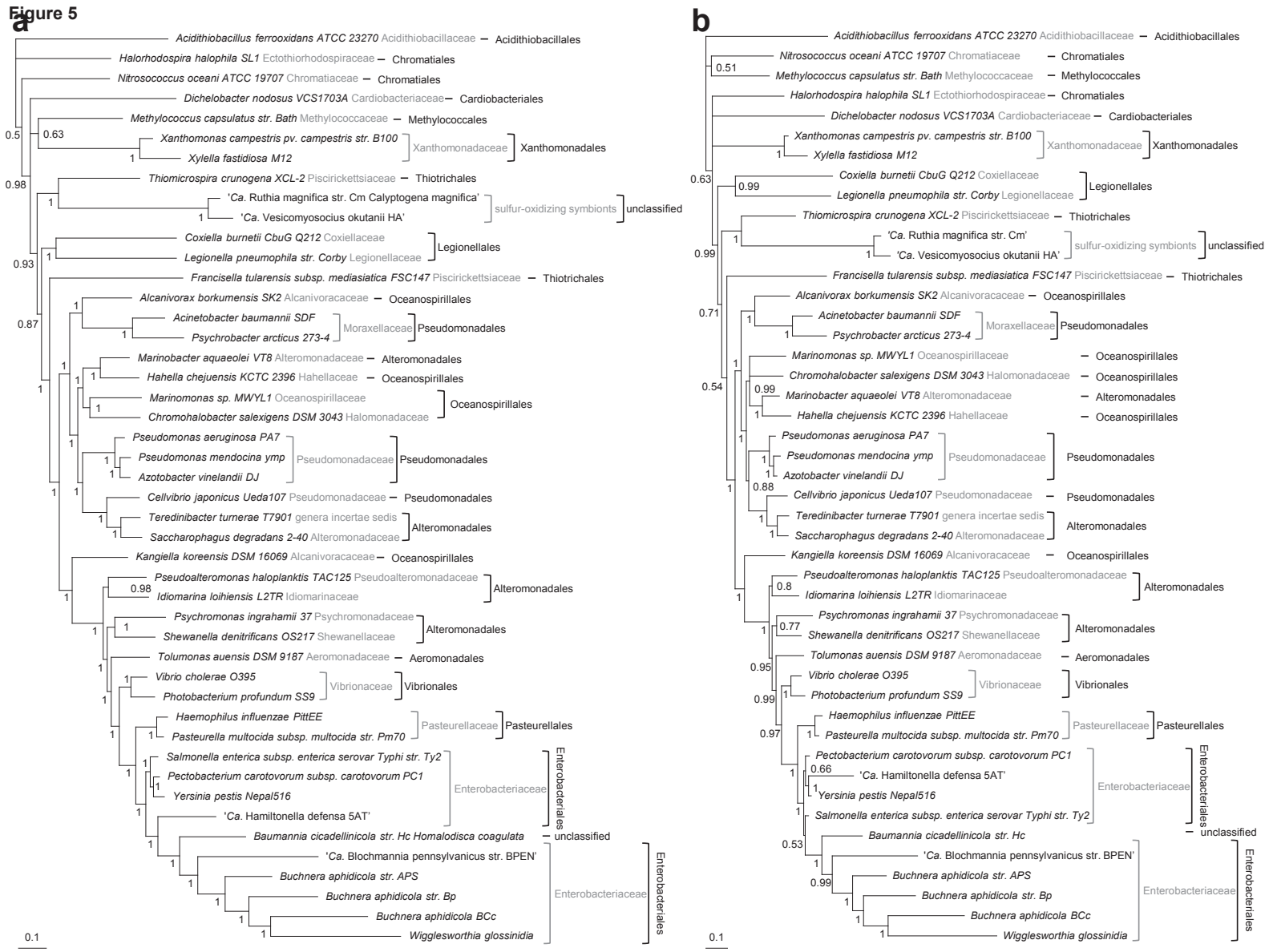

a
Acidithiobacillus ferrooxidans ATCC 23270 Acidithiobacillaceae – Acidithiobacillales
Halorhodospira halophila SL1 Ectothiorhodospiraceae – Chromatiales
Nitrosococcus oceani ATCC 19707 Chromatiaceae – Chromatiales
Dichelobacter nodosus VCS1703A Cardiobacteriaceae – Cardiobacteriales
Methylococcus capsulatus str. Bath Methylococcaceae – Methylococcales
Xanthomonas campestris pv. campestris str. B100
Xylella fastidiosa M12
Xanthomonadaceae
Xanthomonadales
Thiomicrospira crunogena XCL-2 Piscirickettsiaceae – Thiotrichales
'Ca. Ruthia magnifica str. Cm Calyptogena magnifica'
'Ca. Vesicomyosocius okutanii HA'
sulfur-oxidizing symbionts
unclassified
Coxiella burnetii CbuG Q212 Coxiellaceae
Legionella pneumophila str. Corby Legionellaceae
Legionellales
Francisella tularensis subsp. mediasiatica FSC147 Piscirickettsiaceae – Thiotrichales
Alcanivorax borkumensis SK2 Alcanivoracaceae – Oceanospirillales
Acinetobacter baumannii SDF
Psychrobacter arcticus 273-4
Moraxellaceae
Pseudomonadales
Marinobacter aquaeolei VT8 Alteromonadaceae – Alteromonadales
Hahella chejuensis KCTC 2396 Hahellaceae – Oceanospirillales
Marinomonas sp. MWYL1 Oceanospirillaceae
Chromohalobacter salexigens DSM 3043 Halomonadaceae
Oceanospirillales
Pseudomonas aeruginosa PA7
Pseudomonas mendocina ymp
Azotobacter vinelandii DJ
Pseudomonadaceae
Pseudomonadales
Cellvibrio japonicus Ueda107 Pseudomonadaceae – Pseudomonadales
Teredinibacter turnerae T7901 genera incertae sedis
Saccharophagus degradans 2-40 Alteromonadaceae
Alteromonadales
Kangiella koreensis DSM 16069 Alcanivoracaceae – Oceanospirillales
Pseudoalteromonas haloplanktis TAC125 Pseudoalteromonadaceae
Idiomarina loihiensis L2TR Idiomarinaceae
Alteromonadales
Psychromonas ingrahamii 37 Psychromonadaceae
Shewanella denitrificans OS217 Shewanellaceae
Alteromonadales
Tolumonas auensis DSM 9187 Aeromonadaceae – Aeromonadales
Vibrio cholerae O395
Photobacterium profundum SS9
Vibrionaceae
Vibrionales
Haemophilus influenzae PittEE
Pasteurella multocida subsp. multocida str. Pm70
Pasteurellaceae
Pasteurellales
Salmonella enterica subsp. enterica serovar Typhi str. Ty2
Pectobacterium carotovorum subsp. carotovorum PC1
Yersinia pestis Nepal516
'Ca. Hamiltonella defensa 5AT'
Enterobacteriaceae
Enterobacteriales
Baumannia cicadellinicola str. Hc Homalodisca coagulata – unclassified
'Ca. Blochmannia pennsylvanicus str. BPEN'
Buchnera aphidicola str. APS
Buchnera aphidicola str. Bp
Buchnera aphidicola BCc
Wigglesworthia glossinidia
Enterobacteriaceae
Enterobacteriales
0.1
b
Acidithiobacillus ferrooxidans ATCC 23270 Acidithiobacillaceae – Acidithiobacillales
Nitrosococcus oceani ATCC 19707 Chromatiaceae – Chromatiales
Methylococcus capsulatus str. Bath Methylococcaceae – Methylococcales
Halorhodospira halophila SL1 Ectothiorhodospiraceae – Chromatiales
Dichelobacter nodosus VCS1703A Cardiobacteriaceae – Cardiobacteriales
Xanthomonas campestris pv. campestris str. B100
Xylella fastidiosa M12
Xanthomonadaceae
Xanthomonadales
Coxiella burnetii CbuG Q212 Coxiellaceae
Legionella pneumophila str. Corby Legionellaceae
Legionellales
Thiomicrospira crunogena XCL-2 Piscirickettsiaceae – Thiotrichales
'Ca. Ruthia magnifica str. Cm'
'Ca. Vesicomyosocius okutanii HA'
sulfur-oxidizing symbionts
unclassified
Francisella tularensis subsp. mediasiatica FSC147 Piscirickettsiaceae – Thiotrichales
Alcanivorax borkumensis SK2 Alcanivoracaceae – Oceanospirillales
Acinetobacter baumannii SDF
Psychrobacter arcticus 273-4
Moraxellaceae
Pseudomonadales
Marinomonas sp. MWYL1 Oceanospirillaceae – Oceanospirillales
Chromohalobacter salexigens DSM 3043 Halomonadaceae – Oceanospirillales
Marinobacter aquaeolei VT8 Alteromonadaceae – Alteromonadales
Hahella chejuensis KCTC 2396 Hahellaceae – Oceanospirillales
Pseudomonas aeruginosa PA7
Pseudomonas mendocina ymp
Azotobacter vinelandii DJ
Pseudomonadaceae
Pseudomonadales
Cellvibrio japonicus Ueda107 Pseudomonadaceae – Pseudomonadales
Teredinibacter turnerae T7901 genera incertae sedis
Saccharophagus degradans 2-40 Alteromonadaceae
Alteromonadales
Kangiella koreensis DSM 16069 Alcanivoracaceae – Oceanospirillales
Pseudoalteromonas haloplanktis TAC125 Pseudoalteromonadaceae
Idiomarina loihiensis L2TR Idiomarinaceae
Alteromonadales
Psychromonas ingrahamii 37 Psychromonadaceae
Shewanella denitrificans OS217 Shewanellaceae
Alteromonadales
Tolumonas auensis DSM 9187 Aeromonadaceae – Aeromonadales
Vibrio cholerae O395
Photobacterium profundum SS9
Vibrionaceae
Vibrionales
Haemophilus influenzae PittEE
Pasteurella multocida subsp. multocida str. Pm70
Pasteurellaceae
Pasteurellales
Pectobacterium carotovorum subsp. carotovorum PC1
'Ca. Hamiltonella defensa 5AT'
Yersinia pestis Nepal516
Salmonella enterica subsp. enterica serovar Typhi str. Ty2
Enterobacteriaceae
Enterobacteriales
unclassified
Baumannia cicadellinicola str. Hc
'Ca. Blochmannia pennsylvanicus str. BPEN'
Buchnera aphidicola str. APS
Buchnera aphidicola str. Bp
Buchnera aphidicola BCc
Wigglesworthia glossinidia
Enterobacteriaceae
Enterobacteriales
0.1

Figure 6

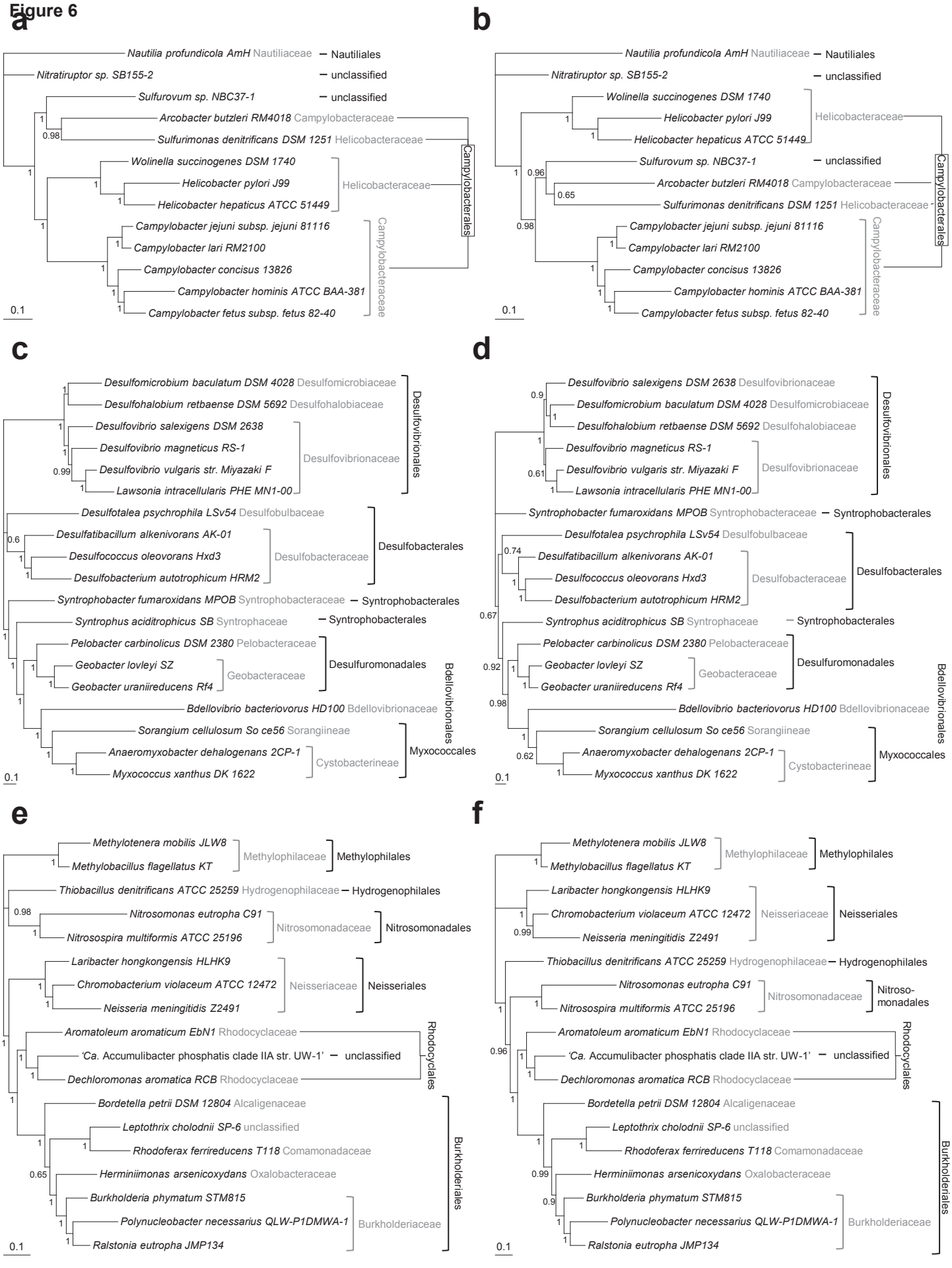
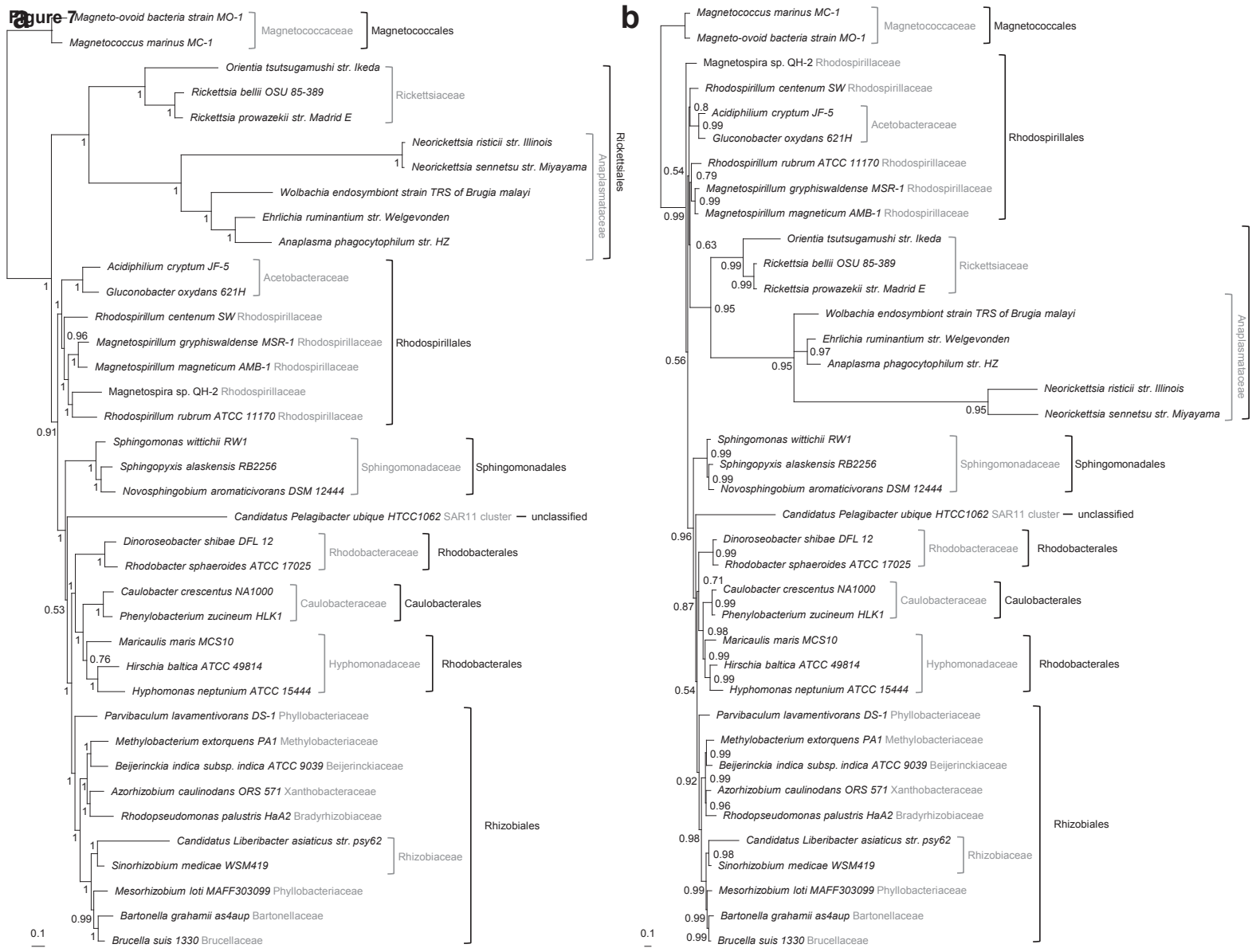

a
Nautilia profundicola AmH Nautiliaceae
Nautiliales
Nitratiruptor sp. SB155-2
unclassified
Sulfurovum sp. NBC37-1
unclassified
Arcobacter butzleri RM4018 Campylobacteraceae
Sulfurimonas denitrificans DSM 1251 Helicobacteraceae
Wolinella succinogenes DSM 1740
Helicobacter pylori J99
Helicobacter hepaticus ATCC 51449
Helicobacteraceae
Campylobacterales
Campylobacter jejuni subsp. jejuni 81116
Campylobacter lari RM2100
Campylobacter concisus 13826
Campylobacter hominis ATCC BAA-381
Campylobacter fetus subsp. fetus 82-40
Campylobacteraceae
0.1
b
Nautilia profundicola AmH Nautiliaceae
Nautiliales
Nitratiruptor sp. SB155-2
unclassified
Wolinella succinogenes DSM 1740
Helicobacter pylori J99
Helicobacter hepaticus ATCC 51449
Helicobacteraceae
Sulfurovum sp. NBC37-1
unclassified
Arcobacter butzleri RM4018 Campylobacteraceae
Sulfurimonas denitrificans DSM 1251 Helicobacteraceae
Campylobacterales
Campylobacter jejuni subsp. jejuni 81116
Campylobacter lari RM2100
Campylobacter concisus 13826
Campylobacter hominis ATCC BAA-381
Campylobacter fetus subsp. fetus 82-40
Campylobacteraceae
0.1
c
Desulfomicrobium baculatum DSM 4028 Desulfomicrobiaceae
Desulfohalobium retbaense DSM 5692 Desulfohalobiaceae
Desulfovibrionales
Desulfovibrio salexigens DSM 2638
Desulfovibrio magneticus RS-1
Desulfovibrio vulgaris str. Miyazaki F
Lawsonia intracellularis PHE MN1-00
Desulfovibrionaceae
Desulfotalea psychrophila LSv54 Desulfobulbaceae
Desulfatibacillum alkenivorans AK-01
Desulfococcus oleovorans Hxd3
Desulfobacterium autotrophicum HRM2
Desulfobacteraceae
Desulfobacterales
Syntrophobacter fumaroxidans MPOB Syntrophobacteraceae
Syntrophobacterales
Syntrophus aciditrophicus SB Syntrophaceae
Syntrophobacterales
Pelobacter carbinolicus DSM 2380 Pelobacteraceae
Geobacter lovleyi SZ
Geobacter uraniireducens Rf4
Geobacteraceae
Desulfuromonadales
Bdellovibrionales
Bdellovibrio bacteriovorus HD100 Bdellovibrionaceae
Sorangium cellulosum So ce56 Sorangiineae
Anaeromyxobacter dehalogenans 2CP-1
Myxococcus xanthus DK 1622
Cystobacterineae
Myxococcales
0.1
d
Desulfovibrio salexigens DSM 2638 Desulfovibrionaceae
Desulfomicrobium baculatum DSM 4028 Desulfomicrobiaceae
Desulfohalobium retbaense DSM 5692 Desulfohalobiaceae
Desulfovibrionales
Desulfovibrio magneticus RS-1
Desulfovibrio vulgaris str. Miyazaki F
Lawsonia intracellularis PHE MN1-00
Desulfovibrionaceae
Syntrophobacter fumaroxidans MPOB Syntrophobacteraceae
Syntrophobacterales
Desulfotalea psychrophila LSv54 Desulfobulbaceae
Desulfatibacillum alkenivorans AK-01
Desulfococcus oleovorans Hxd3
Desulfobacterium autotrophicum HRM2
Desulfobacteraceae
Desulfobacterales
Syntrophus aciditrophicus SB Syntrophaceae
Syntrophobacterales
Pelobacter carbinolicus DSM 2380 Pelobacteraceae
Geobacter lovleyi SZ
Geobacter uraniireducens Rf4
Geobacteraceae
Desulfuromonadales
Bdellovibrionales
Bdellovibrio bacteriovorus HD100 Bdellovibrionaceae
Sorangium cellulosum So ce56 Sorangiineae
Anaeromyxobacter dehalogenans 2CP-1
Myxococcus xanthus DK 1622
Cystobacterineae
Myxococcales
0.1
e
Methylotenera mobilis JLW8
Methylobacillus flagellatus KT
Methylophilaceae
Methylophilales
Thiobacillus denitrificans ATCC 25259 Hydrogenophilaceae
Hydrogenophilales
Nitrosomonas eutropha C91
Nitrosospira multiformis ATCC 25196
Nitrosomonadaceae
Nitrosomonadales
Laribacter hongkongensis HLHK9
Chromobacterium violaceum ATCC 12472
Neisseria meningitidis Z2491
Neisseriaceae
Neisseriales
Aromatoleum aromaticum EbN1 Rhodocyclaceae
'Ca. Accumulibacter phosphatis clade IIA str. UW-1'
unclassified
Dechloromonas aromatica RCB Rhodocyclaceae
Rhodocyclales
Bordetella petrii DSM 12804 Alcaligenaceae
Leptothrix cholodnii SP-6 unclassified
Rhodoferax ferrireducens T118 Comamonadaceae
Herminiimonas arsenicoxydans Oxalobacteraceae
Burkholderia phymatum STM815
Polynucleobacter necessarius QLW-P1DMWA-1
Ralstonia eutropha JMP134
Burkholderiaceae
Burkholderiales
0.1
f
Methylotenera mobilis JLW8
Methylobacillus flagellatus KT
Methylophilaceae
Methylophilales
Laribacter hongkongensis HLHK9
Chromobacterium violaceum ATCC 12472
Neisseria meningitidis Z2491
Neisseriaceae
Neisseriales
Thiobacillus denitrificans ATCC 25259 Hydrogenophilaceae
Hydrogenophilales
Nitrosomonas eutropha C91
Nitrosospira multiformis ATCC 25196
Nitrosomonadaceae
Nitroso-
monadales
Aromatoleum aromaticum EbN1 Rhodocyclaceae
'Ca. Accumulibacter phosphatis clade IIA str. UW-1'
unclassified
Dechloromonas aromatica RCB Rhodocyclaceae
Rhodocyclales
Bordetella petrii DSM 12804 Alcaligenaceae
Leptothrix cholodnii SP-6 unclassified
Rhodoferax ferrireducens T118 Comamonadaceae
Herminiimonas arsenicoxydans Oxalobacteraceae
Burkholderia phymatum STM815
Polynucleobacter necessarius QLW-P1DMWA-1
Ralstonia eutropha JMP134
Burkholderiaceae
Burkholderiales
0.1

Figure 7


a
Magneto-ovoid bacteria strain MO-1
Magnetococcus marinus MC-1
Magnetococcaceae
Magnetococcales
Orientia tsutsugamushi str. Ikeda
Rickettsia bellii OSU 85-389
Rickettsia prowazekii str. Madrid E
Rickettsiaceae
Neorickettsia risticii str. Illinois
Neorickettsia sennetsu str. Miyayama
Wolbachia endosymbiont strain TRS of Brugia malayi
Ehrlichia ruminantium str. Welgevonden
Anaplasma phagocytophilum str. HZ
Anaplasmataceae
Rickettsiales
Acidiphilium cryptum JF-5
Gluconobacter oxydans 621H
Acetobacteraceae
Rhodospirillum centenum SW Rhodospirillaceae
Magnetospirillum gryphiswaldense MSR-1 Rhodospirillaceae
Magnetospirillum magneticum AMB-1 Rhodospirillaceae
Magnetospira sp. QH-2 Rhodospirillaceae
Rhodospirillum rubrum ATCC 11170 Rhodospirillaceae
Rhodospirillales
Sphingomonas wittichii RW1
Sphingopyxis alaskensis RB2256
Novosphingobium aromaticivorans DSM 12444
Sphingomonadaceae
Sphingomonadales
Candidatus Pelagibacter ubique HTCC1062 SAR11 cluster
unclassified
Dinoroseobacter shibae DFL 12
Rhodobacter sphaeroides ATCC 17025
Rhodobacteraceae
Rhodobacterales
Caulobacter crescentus NA1000
Phenylobacterium zucineum HLK1
Caulobacteraceae
Caulobacterales
Maricaulis maris MCS10
Hirschia baltica ATCC 49814
Hyphomonadaceae
Rhodobacterales
Hyphomonas neptunium ATCC 15444
Parvibaculum lavamentivorans DS-1 Phyllobacteriaceae
Methylobacterium extorquens PA1 Methylobacteriaceae
Beijerinckia indica subsp. indica ATCC 9039 Beijerinckiaceae
Azorhizobium caulinodans ORS 571 Xanthobacteraceae
Rhodopseudomonas palustris HaA2 Bradyrhizobiaceae
Candidatus Liberibacter asiaticus str. psy62
Sinorhizobium medicae WSM419
Rhizobiaceae
Rhizobiales
Mesorhizobium loti MAFF303099 Phyllobacteriaceae
Bartonella grahamii as4aup Bartonellaceae
Brucella suis 1330 Brucellaceae
0.1
b
Magnetococcus marinus MC-1
Magneto-ovoid bacteria strain MO-1
Magnetococcaceae
Magnetococcales
Magnetospira sp. QH-2 Rhodospirillaceae
Rhodospirillum centenum SW Rhodospirillaceae
Acidiphilium cryptum JF-5
Gluconobacter oxydans 621H
Acetobacteraceae
Rhodospirillum rubrum ATCC 11170 Rhodospirillaceae
Magnetospirillum gryphiswaldense MSR-1 Rhodospirillaceae
Magnetospirillum magneticum AMB-1 Rhodospirillaceae
Rhodospirillales
Orientia tsutsugamushi str. Ikeda
Rickettsia bellii OSU 85-389
Rickettsia prowazekii str. Madrid E
Rickettsiaceae
Wolbachia endosymbiont strain TRS of Brugia malayi
Ehrlichia ruminantium str. Welgevonden
Anaplasma phagocytophilum str. HZ
Neorickettsia risticii str. Illinois
Neorickettsia sennetsu str. Miyayama
Anaplasmataceae
Rickettsiales
Sphingomonas wittichii RW1
Sphingopyxis alaskensis RB2256
Novosphingobium aromaticivorans DSM 12444
Sphingomonadaceae
Sphingomonadales
Candidatus Pelagibacter ubique HTCC1062 SAR11 cluster
unclassified
Dinoroseobacter shibae DFL 12
Rhodobacter sphaeroides ATCC 17025
Rhodobacteraceae
Rhodobacterales
Caulobacter crescentus NA1000
Phenylobacterium zucineum HLK1
Caulobacteraceae
Caulobacterales
Maricaulis maris MCS10
Hirschia baltica ATCC 49814
Hyphomonadaceae
Rhodobacterales
Hyphomonas neptunium ATCC 15444
Parvibaculum lavamentivorans DS-1 Phyllobacteriaceae
Methylobacterium extorquens PA1 Methylobacteriaceae
Beijerinckia indica subsp. indica ATCC 9039 Beijerinckiaceae
Azorhizobium caulinodans ORS 571 Xanthobacteraceae
Rhodopseudomonas palustris HaA2 Bradyrhizobiaceae
Candidatus Liberibacter asiaticus str. psy62
Sinorhizobium medicae WSM419
Rhizobiaceae
Rhizobiales
Mesorhizobium loti MAFF303099 Phyllobacteriaceae
Bartonella grahamii as4aup Bartonellaceae
Brucella suis 1330 Brucellaceae
0.1

# 4

# La résurrection de protéines ancestrales

## 4.1 Meilleurs modèles, meilleures résurrections.

### 4.1.1 Introduction

Zuckerkandl et Pauling sont souvent considérés comme des pionniers de la phylogénie et de l'évolution moléculaire. Ils sont particulièrement connus pour avoir, dès 1962, proposé le concept d'horloge moléculaire dans un papier mettant en relation le nombre moyen de substitutions d'acides aminés et les temps de divergence entre espèces (Zuckerkandl and Pauling, 1962). Cette observation servira de pilier à la théorie neutre de l'évolution moléculaire formulée plus tard par Kimuara (Kimura, 1968). Par la suite, Pauling and Zuckerkandl (1963) ont publié un autre article dont le titre est très évocateur : 'Chemical Paleogenetics - Molecular "Restoration Studies" of Extinct Forms of Life'. Dès 1963, Pauling and Zuckerkandl ont posé les bases d'un champ de recherche de la biologie évolutive en plein essor actuellement : la résurrection de protéines ancestrales. En guise d'introduction à l'article qui suit et qui traite de l'importance de l'utilisation des modèles hétérogènes et de l'utilisation d'un arbre réconcilié entre l'arbre des séquences et l'arbre des espèces pour réaliser les inférences de séquences ancestrales, je cite ici un passage de l'article de Pauling and Zuckerkandl (1963), montrant à quel point ces deux chercheurs furent

de grands visionnaires :

*Fossil remains no doubt express the activity of only a fraction of the genes of a given organism (although perhaps a significant fraction) and this fraction cannot be analyzed into its components.* ***Paleobiochemistry****, through molecular restoration studies on the basis of existing related polypeptide chains, provides the means of investigating the structure of such components for any part of the genome of extinct organisms. This holds, however, only in relation to structural genes, as long as the object of such studies is confined to the polypeptide products rather than extended to the genic material itself. Yet, one the structures of ancestral polypeptide chains are known, it will in the future be possible to synthesize these presumed components of extinct organisms. Thus one will be able to study the physico-chemical properties of these molecules and to make inferences about their functions*.

L'article qui suit n'est pas encore soumis.

### 4.1.2 Manuscrit

# Biologically motivated models strongly improve the functionality of resurrected proteins

Mathieu Groussin[1], Joanne K Hobbs[2], Gergely J Szöllősi[1,3],
Simonetta Gribaldo[4], Vickery L. Arcus[2], and Manolo Gouy[1]

1 : *Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, Université Lyon 1, CNRS, UMR5558, Villeurbanne, France*

2 : *Department of Biological Sciences, University of Waikato, Hamilton, New Zealand*

3 : *ELTE-MTA "Lendület" Biophysics Research Group, Pázmány P. stny. 1A., H-1117 Budapest, Hungary*

4 : *Institut Pasteur, Département de Microbiologie, Unité de Biologie Moléculaire du Gène chez les Extrêmophiles, 25-28 rue du Dr Roux, 75724 Paris cedex 15, France*

# Abstract

### Background

The resurrection of ancestral proteins holds great potential for both fundamental and applied fields of biology, such as biochemistry, ecology or molecular evolution. By tracing substitutions along a gene phylogeny, ancestral proteins can be reconstructed in silico and subsequently synthesized in vitro. This elegant strategy directly reveals the complex mechanisms responsible for the evolution of protein functions and structures. The overwhelming majority of studies employing ancestral sequence reconstruction (ASR) to date have, however, used simplistic methods, which may have lead to inacurrate biological conclusions.

### Methodology/Principal Findings

ASR requires an evolutionary model and a phylogenetic tree describing the substitution processes and the pattern of descent that produced the sequences under study. Despite evidence that the substitution process is heterogeneous among sites, and that phylogenetic reconstruction is affected by duplications, horizontal transfers and losses of genes (DTL), neither have been considered in resurrection studies. Here, we perform simulations to show that heterogeneous substitution models infer more accurate ancestral sequences, observing a strong correlation between model fit and ASR accuracy. We also find that modeling duplications, horizontal transfers and losses during gene tree reconstruction further increases ASR accuracy, underscoring the importance of tree topology in the inference of putative ancestors. In silico results are validated with in vitro resurrections of the LeuB enzyme for the ancestor of the Firmicutes, which demonstrate that using heterogeneous models and DTL information results in biochemically more realistic and kinetically more stable proteins.

### Conclusions/Significance

As simplistic approaches readily produce functional resurrected protein ancestors, biological conclusions strongly depends on the accuracy of ASR methods. Here, we propose a new protocol for ASR that accounts for the heterogeneity of the substitution process and gene level events (DTL). This protocol should be used in future protein resurrection studies to accurately decipher how natural selection has shaped the proteins of today.

# Introduction

Ancestral sequence resurrection allows to gain insights into the evolutionary processes that have shaped the structure and function of extant proteins by studying the properties of their now extinct ancestors (Chang and Donoghue, 2000; Harms and Thornton, 2010). 50 years ago Pauling and Zuckerkandl (1963) proposed that the resurrection of ancestral sequences inferred *in silico* could open the possibility of experimentally studying the ancestors of modern proteins. This is possible because, given a set of homologous sequences, a corresponding phylogenetic tree, and a model of sequence evolution, one can infer ancestral sequences for any node of the phylogeny. These putative ancestral sequences can then be "resurrected" in the laboratory using standard molecular biology techniques, giving access to extinct proteins and their phenotypes. Since the work of Malcolm et al. (1990) and Stackhouse et al. (1990), who first implemented this idea in practice, numerous studies combining ancestral sequence reconstruction (ASR) with experimental resurrection have investigated diverse biological questions, ranging from ancient adaptations to temperature (Gaucher et al., 2003, 2008; Hobbs et al., 2012), to ancestral ecological adaptations (Chang et al., 2002; Mirceta et al., 2013), to the emergence of protein function (Benner et al., 2002; Ortlund et al., 2007), to the influence of gene duplication on functional divergence (Voordeckers et al., 2012), to the evolution of molecular complexes (Finnigan et al., 2012), to industrial or biomedical applications (Kodra et al., 2007; Cole and Gaucher, 2011).

With the increase in popularity of the ASR approach, several methodological improvements have been proposed (Yang et al., 1995; Koshi and Goldstein, 1996; Pupko et al., 2000; Williams et al., 2006; Pupko et al., 2007). The parsimony approach of (Fitch, 1971) used in early studies (Jermann et al., 1995) has been supplanted by Maximum Likelihood (ML) (Yang et al., 1995; Pupko et al., 2000), which has the advantage of providing a measure for uncertainty of the reconstruction and of allowing the development of more elaborate substitution models that more fully capture the complexity of the underlying evolutionary process. Using ML, Yang et al. (1995) proposed the marginal reconstruction algorithm that we considered in this study and which is used in almost all modern ASR studies. This approach allows us to compute the likelihood of each possible ancestral state for each internal node at each site in the sequence alignment. The state having the highest likelihood is considered as the putative ancestral state. Posterior probabilities are then computed for each possible state, providing confidence in the reconstruction inference (Yang et al., 1995). Despite the flexibility afforded by such a probabilistic approach, and the correspondingly wide range of available substitution models and tree reconstruction algorithms, few studies have focused on the effect of the substitution model or of the phylogenetic tree on ASR.

In models of sequence evolution it is usually assumed that all residues of a protein evolve according to a constant substitution process. In this case, the Markovian substitution model specifying substitution

rates between amino-acids is said to be homogeneous among sites. However, if all homologous sites of a protein alignment evolved according to the same process, they would display homogeneous amino acid frequencies along the complete sequence. Due to functional and structural selective constraints acting on native proteins, biological sequences do not display this property (Koshi and Goldstein, 1998; Lartillot and Philippe, 2004). Elaborate amino acid substitution models have therefore been developed to capture the variation of the process among sites, and have been shown to more accurately fit biological sequence data (Le et al., 2008b). Such models are called site-heterogeneous. Given their better fit to data, these models can be expected to yield more accurate ancestral sequences. However, despite the availability of these models in the literature as well in publicly available software libraries, no studies attempting to perform ASR on protein sequences of interest have used them, relying rather on relatively simple site-homogeneous substitution models such as JTT (Jones et al., 1992), WAG (Whelan and Goldman, 2001) or LG (Le and Gascuel, 2008).

In addition, we are also concerned by the second major component used in all ASR studies, the phylogenetic tree along which ancestral sequences are inferred (the gene tree). In most, if not all, previous studies where ASR and resurrection were performed, ancestral sequences were inferred using a gene tree reconstructed using only the multiple sequence alignment of existing sequences (Harms and Thornton, 2010); we refer to such gene trees as *sequence-only trees*. Unfortunately, individual sequences alone contain limited signal, and as a result phylogenetic reconstruction almost always involves choosing between statistically equivalent or weakly distinguishable relationships. Furthermore, while each set of homologous genes has its own unique story, they are all related by a shared species history, which could be helpful for gene tree inference. To exploit this possibility, genome evolutionary processes such as duplication, horizontal transfer and loss must be modeled to *reconcile* the gene tree with the species tree (Szöllősi et al., 2012). The advantage of such "species tree aware" methods is that they allow us to detect and correct tree reconstruction errors resulting from the finite size of alignments or the inadequacy of the substitution model employed, while at the same time retaining *bona fide* phylogenetic discord produced by genome evolutionary proceses. Methods that combine the substitution model with models of genome evolution to reconstruct *joint trees* have demonstrated that taking into account information on the species tree can dramatically increase the accuracy of gene trees (Åkerborg et al., 2009; Rasmussen and Kellis, 2012; Wu et al., 2013; Boussau et al., 2013; Szöllősi et al., 2013a) (Figure 1).

The purpose of this study is to investigate to what extent ASR can benefit from the use of such biologically realistic models of substitution and tree reconstruction.

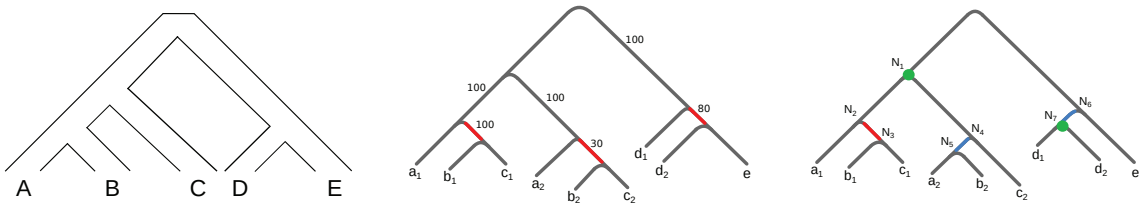



Figure 1: **Gene tree/species tree reconciliation**. We propose here a simple example highlighting the impact of gene tree/species tree reconciliation on the topology of the gene tree along which ASR may be performed. In this example, a gene family is supposed to evolve along the species tree in a). In b), the phylogenetic tree reconstructed only with sequence information is represented. Two copies of this gene are present in species A, B, C and D. Values on branches represent branch support, which can be any type of support. Branches in red are bipartitions that are inconsistent with the species tree. In c), the reconciled joint tree between sequence and species information is represented and can be obtained with a model that jointly calculates the sequence and the genomic processes (such as duplications/transfers/losses) likelihoods. On the left part of the tree, the inconsistency is strongly supported by the sequence information and conserved in the reconciled tree. An horizontal gene transfer can be assumed between species B and C. In the middle part of the tree, there is not enough support for a possible transfer (the support is 30/100), so that the inconsistency is corrected (blue branch). The model also inferred the presence of a deep duplication, prior to the emergence of species A, B and C. In the right part of the tree, the inconsistency is also corrected. In this example, we assumed that sequence information does not support enough the close relationship between d2 and e, which would have to be explained by a duplication prior to the divergence of D and E, followed by a loss of one copy in E. Instead, it is more likely to infer a single duplication in the branch leading to D.

# Material & Methods

## Data used for *in silico* experiments

To perform *in silico* experiments to investigate the influence of the substitution model and the phylogenetic tree on ASR, we used the dataset from Szöllősi et al. (2012). This dataset comprises 1,099 gene families from 36 cyanobacterial genomes available in the HOGENOM database (Penel et al., 2009). The phylogenomic species tree of these 36 species that Szöllősi et al. (2012) reconstructed was also used in the present study. With this species topology and a newly describred model of gene tree/species tree reconciliation, Szöllősi et al. (2013b) computed the reconciled joint trees for the 1,099 families. Here, we randomly chose 100 families out the 1,099 and we simulated sequences along their corresponding joint tree topologies, that we considered as 'true' gene trees. We added an outgroup species to both the species tree and 'true' topologies. The branch length leading to the outgroup species was set to 1/2 of the joint tree height.

## Substitution models employed in this study

All models employed in this study are empirical Markovian substitution models, in which all parameters were previously learned on a large database and fixed to be subsequently used on other datasets. Furthermore, they were all used in combination with a discrete $\Gamma$ distribution to model the site-specific

rate variation, with four categories.

Site-homogeneous substitution models assume that the evolutionary process is constant among sites. When the Markovian process is time-reversible, the transition probability matrix is computed by multiplying the matrix of exchangeabilities with the diagonal matrix of equilibrium frequencies (Whelan and Goldman, 2001). Hereafter, the latter matrix is refered to as a profile of equilibrium frequencies. In the case of site-homogeneous models, both the exchangeabilities and the profile are constant among sites. The JTT and LG site-homogeneous models were employed.

To model the heterogeneity of the process among sites, mixture models were considered. These approaches use sets of different models in which each model is assigned a particular weight. The likelihood of a given site is then the sum of all weighted likelihoods computed with each model of the mixture (Le et al., 2008b). The models of the mixture may have been learned to take into account protein properties that are heterogeneous along the sequence and that influence the substitution process, such as solvent exposure or secondary structure. In line with this, Le et al. (2008b) and Le and Gascuel (2010) proposed a series of empirical mixture models that outperform any site-homogeneous models. They learned their models on the HSSP database (Schneider et al., 1997) of aligned protein sequences (Schneider et al., 1997) in a supervised or unsupervised way. In the supervised way, sites were *a priori* assigned to a component of the mixture given knowledge about their localization in the protein, and the exhangeabilities and equilibrium frequencies of each model were subsequently learned from these sites. Le et al. (2008b) and Le and Gascuel (2010) inferred four models in this way:

- EX2, which is composed of two matrices corresponding to exposed/buried sites
- EX3, which is composed of three matrices corresponding to highly exposed/intermediate/buried sites
- EHO, which is composed of three matrices corresponding to extended/helix/other sites
- EX_EHO, which is composed of six matrices corresponding to the combination of EX2 and EHO.

In the unsupervised way, both site partitions and their corresponding matrices were directly learned from the data. Two models were proposed by Le et al. (2008b):

- UL2, which is composed of two matrices
- UL3, which is composed of three matrices

Note that all these models are mixtures of matrices with both exchangeabilities and equilibrium frequencies varying among components.

Mixture of profiles were also previously proposed (Le et al., 2008a). In these site-heterogeneous models, only the profiles vary among the components of the mixture, which share the same exchangeabilities. The components of these mixtures were learned in an unsupervised way. Six models were proposed, with 10, 20, 30, 40, 50 or 60 different profiles and named C10 to C60.

To evaluate the ability of a substitution model to efficiently fit the data, we used the AIC criterion (Akaike, 1973). This criterion allows the evaluation of non-nested models by penalizing the number of parameters influencing the likelihood. The AIC criterion is computed as follows:

$$AIC = -2 \times lnL + 2 \times K,$$

with $lnL$ the final likelihood and $K$ the total number of parameters. For a site-homogeneous model, only the $\alpha$ parameter of the $\Gamma$ distribution is involved, so that $K = 1$. For site-heterogeneous models, the sum of all component-specific weights equals 1, so that $K = (n - 1) + 1$, with $n$ the number of components of the mixture model.

## Simulations

Available substitution models may contain several parameters aiming at capturing molecular footprints left by biological processes during evolution. Even so, they are too simplistic in comparison with the complexity of processes acting on biological data. To mimic this gap between simplicity of substitution models and complexity of biological data, we used a relatively complex model to simulate sequences along the 100 'true' gene trees, and reconstructed phylogenetic trees and ancestral sequences with simpler models. The site-heterogeneous C60 model was used to simulate alignments using the original alignment sizes of the 100 cyanobacterial families. Simulations were performed with our own C++ program depending on Bio++ libraries (Dutheil et al., 2006; Guéguen et al., 2013). For a given alignment, because sites are supposed to evolve independently, all 60 components of the mixture were used to simulate sub-alignments with a number of sites proportional to their empirical weight, with all sub-alignments being subsequently concatenated to produce the final alignment.

## Ancestral Sequence Reconstruction

Given a tree and an alignment, ML estimates of branch lengths and parameters of the substitution models were inferred with bppML, which belongs to the bppSuite of programs (Dutheil and Boussau, 2008) and depends on Bio++ libraries (Guéguen et al., 2013). For all mixture models, the weight of each component was optimized by ML. With these ML estimates, ancestral sequences were then inferred with bppAncestor (Dutheil and Boussau, 2008) using the marginal ASR approach (Yang et al., 1995).

For a given site at a given internal node of the tree, the state having the maximum posterior probability was inferred as the putative ancestral state.

## ASR accuracy measurement

Inferred ancestral sequences were compared to 'true' internal sequences by computing two distances:

- the raw distance, which is simply the number of amino-acid differences divided by the length of the sequence,
- the Miyata distance (Miyata et al., 1979), defined as the amino acid pair distance computed with the Miyata distance matrix, which allows to take into account biochemical similarities between amino-acids in terms of polarity and volume.

## Sequence-only tree/Species tree reconciliations

Szöllősi et al. (2013b) recently described a probabilistic reconciliation model that accounts for the duplication, transfer and loss of genes along a species tree. Given a fixed species tree, the model allows to explore possible paths along which a gene tree may have been generated by a series of speciations, duplications, transfers and losses. To efficiently explore the space of all reconciled trees according to the joint sequence-reconciliation likelihood that combines sequence information and information on the species phylogeny, Szöllősi et al. (2013a) proposed the ALE (Amalgamated likelihood estimation) algorithm. ALE makes use of a sample of gene trees (for instance, a sample of posterior trees produced by a Bayesian program such as Phylobayes (Lartillot et al., 2009)) to compute conditional clade probabilities (Höhna and Drummond, 2012), which are used to approximate the posterior probability of all gene trees that can be amalgamated from clades present in the sample.

ALE was used to perform all sequence-only tree/species tree reconciliations for both simulated and biological (see below) datasets. For each simulated alignment, PhyloBayes (version 3.3f) was run to obtain an MCMC sample of trees using a simple F81 (Poisson) substitution model. Two chains were run in parallel to check for convergence, with a burn-in of 1000 samples followed by at least 10000 samples. These MCMC samples were then used by ALE to explore the space of reconciled trees in combination with the ML estimation of duplication, transfer and loss rates, to eventually propose the *joint tree* , i.e. the reconciled gene tree that maximises the joint sequence-reconciliation likelihood. ALE calculations were performed with the species tree initially used to compute the 'true' gene trees (see above).

## Statistical tests

All statistical tests were performed with R (Team, 2013). The Fisher's combined probability test was realized with the *combine.test* function belonging to the *survcomp* package (Schroeder et al., 2011), available in the Bioconductor set of packages (Gentleman et al., 2004).

## Firmicutes data used for experimental validation

In order to resurrect and experimentally investigate the biochemical properties of the LeuB sequence from the last common ancestor of the Firmicutes, we reconstructed ancestral sequences of the LeuB enzyme along the phylogenetic tree of this bacterial phylum.

### Firmicutes species tree and LeuB sequence-only tree reconstructions

Firmicutes genomic sequences were downloaded from the NCBI, as of April 2012. Orthologous gene families corresponding to all 53 bacterial ribosomal proteins were constructed with Blast. Each individual gene was aligned with Mafft (Katoh and Standley, 2013) and ambiguous sites were trimmed by BMGE (Criscuolo and Gribaldo, 2010), using the BLOSUM30 matrix. Only 46 out of the 53 ribosomal gene alignments were then concatenated. The remaining seven genes (L25, L30, L32, L33, S4, S14, S21) were discarded owing to either the presence of paralogs or a patchy distribution over Firmicutes species. To root both the species tree and the LeuB tree, we incorporated two outgroup LeuB sequences from two Actinobacteria species, *Corynebacterium glutamicum* and *Streptomyces coelicolor*. The final alignment contains 68 Firmicutes species and the species tree was computed with Phylobayes (Lartillot et al., 2009), using the CAT model (Lartillot and Philippe, 2004). Two independent chains were run in parallel to check for convergence. The model of Szöllősi et al. (2013b) used by ALE (Szöllősi et al., 2013a) to search for the joint tree needs divergence times between speciation nodes to compute the probabilities of gene transfers between branches. Therefore, the species tree was calibrated with relative times using Phylobayes and an arbitrary calibration of 1,000 time unit at the root. The Log-normal autocorrelated relaxed clock model (Thorne et al., 1998) was chosen to allow substitution rates to vary in time.

The gene family corresponding to the 71 LeuB sequences found in the 68 species was reconstructed and a preliminary alignment was inferred using Muscle (Edgar, 2004) and used to build a phylogenetic tree using PhyML (Guindon et al., 2010) with the LG model and a $\Gamma$ distribution for rate variation. This preliminary sequence-only tree was used as a guide tree in Prank (Löytynoja and Goldman, 2008) to re-align LeuB sequences. The final LeuB sequence-only tree along which ancestral sequences were reconstructed with PhyloBayes, using the LG+$\Gamma$(4) model, and rooted on the branch between the

Firmicutes and outgroup LeuBs. Three chains were run in parallel to ensure that convergence of the MCMC was reached.

### LeuB joint tree reconstruction

We used the model described in Szöllősi et al. (2013b) and implemented in the ALE program (Szöllősi et al., 2013a) to search for the ML joint reconciled tree, i.e. the reconciled gene tree that maximises the joint sequence-reconciliation likelihood. ALE used the sample of sequence-only trees produced by Phylobayes (see above) and the calibrated species tree to compute the joint tree along which ASR was performed. The joint tree was used as a guide tree in Prank to compute the final alignment in combination with the inference of ancestral gaps, which were subsequently incorporated into ancestral sequences.

### Model selection and fit to the LeuB data

ASR of LeuB was performed both with the site-homogeneous LG substitution model and the site-heterogeneous EX_EHO model. EX_EHO was deemed to be the best site-heterogeneous model at fitting the LeuB data according to the AIC criterion, in comparison with all other site-heterogeneous models tested.

## Typical protocol for ASR

Given the *in silico* and *in vitro* results obtained in the present study, we propose a standard protocol for ASR in ML, which accounts for the use of complex evolutionary models and species tree/gene tree reconciliation. See Supplementary Figure 5 for an illustration of this protocol.

1. Construct the gene family of interest
2. Align homologous sequences and reconstruct the corresponding phylogenetic tree
3. Re-align the sequences with the previous tree used as a guide tree
4. Reconstruct the sequence-only tree
5. Reconstruct the (time-calibrated) species tree
6. Use a species tree/gene tree reconciliation method to reconstruct the reconciled joint tree
7. Re-align the sequences with the reconciled tree used as a guide tree and infer ancestral gap positions

8. Compare site- and time-homogeneous models with more complex models (*i.e.* site- or time-heterogeneous models) with the joint reconciled tree to determine the best model in terms of data-fitting with model selection criteria (LRT, AIC, BIC)

9. Use the parameter estimates of the best substitution model to perform ASR along the joint reconciled tree

10. Incorporate gaps within ancestral sequences

Note that depending on the reconciliation program used, the sequence-only tree used may not consist of a single tree but may consist of a posterior sample of trees (as in this study, with the use of the ALE program). The time-calibration of the species tree is used by reconciliation algorithms to compute proobabilities of lateral gene transfers. Consequently, the species tree need not necessarily be time-calibrated if ASR is performed on a group of species in which lateral gene transfers are thought to be negligible. Step 7 and 8 may be performed jointly with the use of the Prank program (Löytynoja and Goldman, 2008). Currently, only Bio++ libraries (Guéguen et al., 2013) allow the comparison of a large set of homogeneous and heterogeneous amino acid substitution models by ML (with the bppML program (Dutheil and Boussau, 2008)) and the reconstruction of ancestral sequences (with the bppAncestor program (Dutheil and Boussau, 2008)) using these models.

## Biochemical methods

### Protein expression and purification

Gene sequences for the three inferred versions of the ancestral Firmicutes LeuB were codon optimised for expression in *Escherichia coli* and chemically synthesised by Geneart (Life Technologies) with a 5' *NcoI* site and a 3' *PstI* site. Following ligation of the genes into the protein expression vector pPROEX HTb, recombinant proteins were expressed in *E. coli* DH5$\alpha$ with 1 mM IPTG induction at 37°C for 24 hours. Proteins were purified to >= 95% purity by nickel affinity chromatography and subsequent size-exclusion chromatography using the buffers detailed in Hobbs et al. (2012). Protein concentrations were determined using a NanoDrop 2000 (Thermo Scientific) and extinction coefficients calculated using ProtParam on the ExPASy server (web.expasy.org/protparam/).

### LeuB enzyme characterisation

LeuB activity was measured by following the reduction of NAD at 340 nm as described in Hobbs et al. (2012). The $V_{max}$ and Michaelis-Menten constants for both substrates (isopropylmalate; IPM and NAD) were found using the Michaelis-Menten non-linear fitting function in Graphpad Prism 6.

Thermoactivity profiles were determined by measuring the initial rate of activity at $1-5°C$ intervals over a $20-30°C$ temperature range in triplicate. Thermoactivity profile reactions contained 15 mM IPM, 50 mM NAD and 10-50 $\mu$M LeuB enzyme. The free energy of unfolding, $\Delta G^{\ddagger}_{N-U}$, for each enzyme was determined from urea unfolding rates as described in Hobbs et al. (2012).

# Results

## Impact of the substitution model on ASR

### Site-heterogeneous substitution models and ASR accuracy

We first investigated the influence of the substitution model on ASR accuracy. The dataset of Szöllősi et al. (2012) comprises 1099 gene families from 36 cyanobacterial genomes. For each of these real gene families, Szöllősi et al. (2012) computed a reconciled tree. In the present study, we randomly chose 100 families out the 1099 and we simulated sequences along the reconciled tree topologies, considered as 'true' gene trees. ASR was then performed along these 100 'true' gene trees. Here, we focused on the comparison between site-homogeneous and site-heterogeneous models, assuming either homogeneity or heterogeneity of the evolutionary process among sites, respectively. The performance of the LG (Le and Gascuel, 2008) site-homogeneous model was compared to the UL3, C20 and C60 site-heterogeneous models (Le et al., 2008b,a). To evaluate reconstruction accuracy, we measured two distances when comparing inferred ancestral sequences to 'true' sequences recorded during simulations (See Methods). Although the distances between sequences inferred by the two approaches are relatively small, Table 1 shows that, almost systematically, the site-heterogeneous models produce significantly better ancestral sequences. The accuracy of the reconstruction drops linearly with the distance to the leaves, whatever the type of model (for instance, 98% (LG) and 97% (UL3) of Pearson correlation tests are significant ($p-value < 0.05$) after a Bonferroni correction for multiple tests). However, Table 1 and Figure 2 show that site-heterogeneous models produce less reconstruction errors when the distance to the leaves increases (first to fifth quintile) in comparison with LG. Note that all these results are similarly obtained when substituting LG with the site-homogeneous JTT model (Jones et al., 1992).

### Fit to the data and accuracy of the reconstruction

When one attempts to perform ASR on biological data, the ancestors of extant biological sequences are unknown, such that there is no direct possibility to evaluate the accuracy of the reconstruction between models. Therefore, one needs an objective criterion to choose a particular substitution model over others to perform ASR. Although complex evolutionary models, such as site-heterogeneous models,

| Substitution model | All distances to leaves | 1st quintile of distances | 2nd quintile of distances | 3rd quintile of distances | 4th quintile of distances | 5th quintile of distances |
|---|---|---|---|---|---|---|
| LG | 0.067 | 0.003 | 0.018 | 0.049 | 0.091 | 0.149 |
| | 0.093 | 0.004 | 0.026 | 0.068 | 0.126 | 0.209 |
| UL3 | $0.065^{***}$ | $0.003^{NS}$ | $0.017^{**}$ | $0.047^{***}$ | $0.089^{***}$ | $0.146^{***}$ |
| | $0.092^{**}$ | $0.004^{NS}$ | $0.025^{***}$ | $0.067^{*}$ | $0.124^{**}$ | $0.205^{***}$ |
| C20 | $0.057^{***}$ | $0.002^{***}$ | $0.015^{***}$ | $0.040^{***}$ | $0.075^{***}$ | $0.129^{***}$ |
| | $0.081^{***}$ | $0.003^{**}$ | $0.022^{***}$ | $0.056^{***}$ | $0.107^{***}$ | $0.185^{***}$ |
| C60 | $0.055^{***}$ | $0.002^{***}$ | $0.015^{***}$ | $0.039^{***}$ | $0.074^{***}$ | $0.127^{***}$ |
| | $0.080^{***}$ | $0.003^{**}$ | $0.022^{***}$ | $0.055^{***}$ | $0.104^{***}$ | $0.181^{***}$ |

Table 1: **Comparison of ASR accuracy between LG and site-heterogeneous models.** For each model, the first and second row contains the average raw and Miyata distances to 'true' sequences, respectively. Comparison between LG and site-heterogeneous models was measured with paired Student tests, either over all distances to leaves or for each of the five quintiles of distances. $NS$: Non-Significant; $* : p-value < 0.05$; $** : p-value < 0.01$; $*** : p-value < 0.001$.

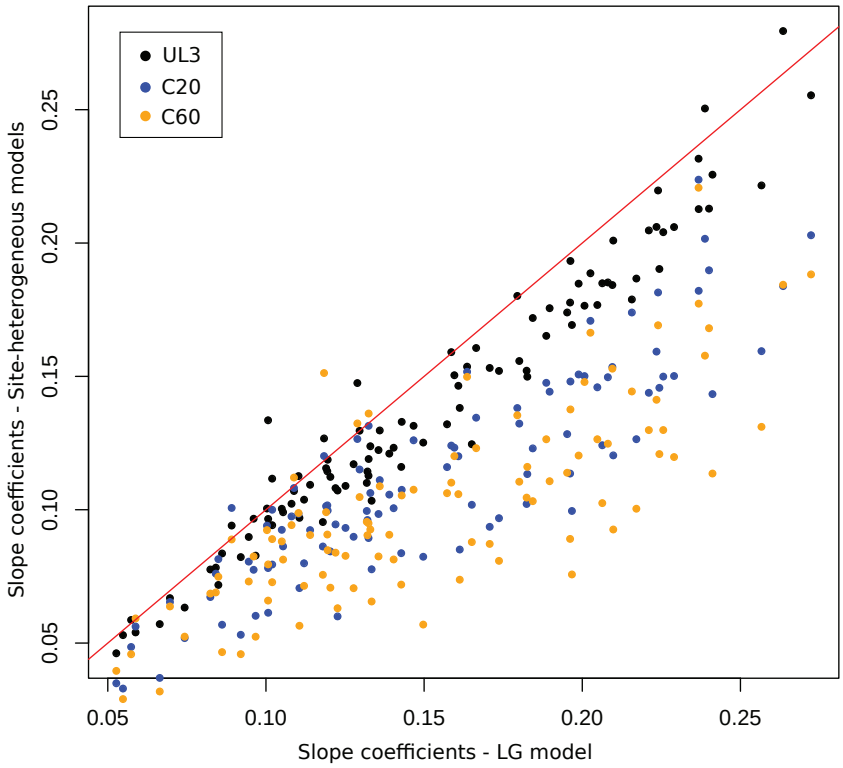



Figure 2: **Site-heterogeneous models infer more accurate ancestral sequences**. For each simulated dataset and for each model, a linear regression forced to pass through the origin was calculated between ASR accuracy measured with the raw distance and the phylogenetic distance to the leaves. For a given dataset, slope coefficients obtained with one of the three site-heterogeneous models ($y$ axis) are compared to those obtained with the site-homogeneous LG model ($x$ axis). The value of a given slope coefficient indicates how inaccurate the model is at inferring ancestral sequences when the distance to the leaves increases. The red line represents the $y = x$ line.

are usually more realistic and provide better likelihoods, they require the estimation of a larger number of parameters. Model selection criteria, such as AIC (Akaike, 1973) aim at balancing the effect of the number of parameters on the final likelihood to select the model that best fits the data. Chang et al.

(2002) and Pupko et al. (2007) have already suggested the use of such criteria to select the model used for ASR. However, no formal evidence has been provided regarding the relationship between fit and ASR accuracy. To do so, a set of site-homogeneous (JTT, LG) and site-heterogeneous (EX2, EX3, EHO, UL2, UL3 citep Le08b, EX_ EHO (Le and Gascuel, 2010), C10 to C60 (Le et al., 2008a); see Material and Methods) substitution models were compared to test whether a correlation exists between the fit of the model to the data and the accuracy of the reconstruction. To do so, the Fisher's combined probability test was used to combine the results from the 100 independent Spearman correlation tests performed on each simulated dataset and test for a correlation between AIC values and average raw distances to 'true' sequences. Supplementary Figure 1 shows the distribution of the 100 $p-values$. The results support that ASR accuracy is strongly and significantly linked to data fitting performance ($p-value < 0.001$, Fisher's combined probability test), underlining the need to test multiple substitution models, including heterogeneous ones, before considering ASR. Note that the test is strongly significant ($p-value < 0.001$) even without the site-heterogeneous mixtures of profiles (models C10 to C60), which are defined in the same way as the model used to simulate sequences (C60).

**Biological data analysis**

As simulations do not reproduce the complexity of biological data, it is worth noting that the weak distances between ancestral sequences inferred by the two types of models should not be considered as predictions of what could be obtained when model comparison is performed on biological alignments. To illustrate this, we inferred ancestral sequences on biological data and compared the results obtained between LG and site-heterogeneous models. To do so, we performed ASR for the 100 cyanobacterial gene families previously considered, along the corresponding 'true' reconciled trees (See above and Material and Methods). On average, the percentage of amino acid differences reaches 2%, 3%, 5% and 5% between the site-homogeneous LG model and the site-heterogeneous EX_EHO, UL3 and both C20 and C60 models respectively, which represents on average about 7 to 20 amino acid differences per ancestral sequence of average length of 375 amino-acids.

Model choice thus has an impact on the accuracy of the most likely ancestral sequences. However, the uncertainty of the reconstruction also needs to be taken into account. Usually, when ancestral states are ambiguously reconstructed, where, for a given residue, several states have posterior probabilities (PP) higher than an *a priori* threshold of 0.2 or 0.3, several versions of the ancestral sequence are reconstructed and experimentally characterized (Finnigan et al., 2012; Voordeckers et al., 2012). By doing so, one can verify the robustness of ancestral functional inferences relatively to the uncertainty of the *in silico* reconstruction. Therefore, one can wonder whether the differences in ancestral amino acid inferences between two models only concern ambiguously reconstructed sites. For residues that are identical between the LG and the C20 models, the average PP is 0.87 with LG and 0.85 with C20.

For residues that are different, the average PP drops to 0.49 with LG and 0.45 with C20, showing that differences in the reconstruction are more concentrated on sites that are intrinsically difficult to reconstruct. However, standard deviations of PP are high (0.21 with LG and 0.19 with C20), so that, with the LG model, up to 12% and 19% of residues that are reconstructed differently by C20 have PP higher than 0.8 and 0.7 respectively, which are usual cutoffs considered in ASR experiments (Konno et al., 2011; Finnigan et al., 2012; Voordeckers et al., 2012) to assume confidence in the ancestral residue.

## Impact of the phylogenetic tree on ASR

To evaluate the impact of using reconciled gene trees that maximise the joint sequence-reconciliation likelihood on the accuracy of ASR, we used the same set of 100 simulated alignments. With these simulated alignments, the corresponding sequence-only trees were reconstructed either with PhyML and the site-homogeneous LG model or with PhyML-CAT (Le et al., 2008a) and the site-heterogeneous C60 model. To compute the joint trees, i.e. reconciled gene trees that maximise the joint sequence-reconciliation likelihood, the ALE program (Szöllősi et al., 2013a) was used (see Material and Methods). Ancestral sequences were then computed along these sequence or joint trees. For nodes defining similar monophyletic clades between the sequence or joint tree and the 'true' tree, these ancestral sequences were compared to the 'true' ancestral sequences recorded during the simulation.

Figure 3a shows that, on average, the sequence trees reconstructed either with LG or with C60 contain more topological errors than the joint trees.

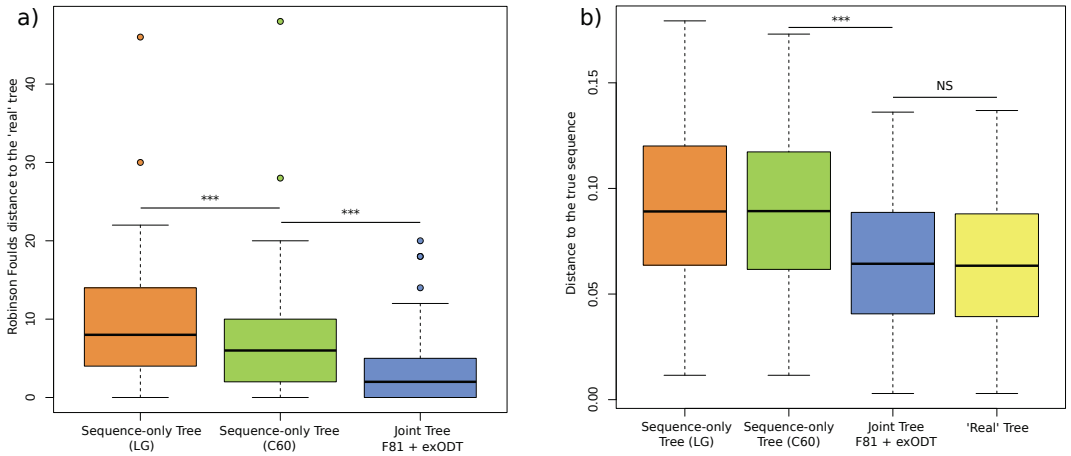



Figure 3: **Impact of the phylogenetic tree on ASR**. a) Phylogenetic reconstruction accuracy. Robinson-Foulds distances were computed between sequence-only trees (LG or C60) or joint trees and the 'true' tree. The exODT model is the reconciliation model described in Szöllősi et al. (2013b) b) ASR accuracy depending on the phylogenetic tree. Distances between inferred and 'true' ancestral sequences were computed for nodes defining similar monophyletic clades between the sequence-only or joint tree and the 'true' tree. ***: $p\text{-}value < 0.001$; NS: non-significant.

These results confirm the findings of Szöllősi et al. (2013a), showing that joint trees are more accurate than sequence-only trees, even when they are reconstructed with the complex model used to simulate

the sequences (C60). Furthermore, this has a direct impact on the ASR accuracy: when ancestral sequences are reconstructed along the joint trees, the accuracy is greatly and significantly improved (Fig. 3b), and is close to the accuracy obtained with the 'true' trees.

We then examined the PP for residues inferred differently with the C60 sequence-only trees and with the joint trees. The average PP reaches 0.82 and 0.81 for the sequence-only trees and joint trees respectively. This shows that the difference in inferences can involve residues that are unambiguously reconstructed with the sequence-only trees, and that the use of joint trees can radically change ancestral predictions and, potentially, subsequent biological conclusions.

## Resurrection and Experimental validation

We have previously used the biochemical and biophysical properties of reconstructed ancestral LeuB enzymes to investigate thermal adaptation in *Bacillus* (Hobbs et al. 2012). Here, we have used the same approach to compare three versions of the same ancestral LeuB enzyme from the last common ancestor of the Firmicutes, the bacterial phylum to which *Bacillus* belongs. These enzymes were inferred and resurrected to investigate the influence of the phylogenetic tree and of the substitution model on potential biological conclusions. The first two enzymes were reconstructed with the LeuB joint tree, with either the site-homogeneous LG model ($\text{LeuB}^{LG}_{joint}$) or the site-heterogeneous EX_EHO model ($\text{LeuB}^{EX\_EHO}_{joint}$). The ALE program, which was used to reconcile sequence and species information, detected 0 duplication, 14 lateral gene transfers and 15 losses. The joint tree has a Robinson-Foulds distance with the sequence-only tree equal to 32, which is very high. The third enzyme was reconstructed with the LeuB sequence-only tree and the EX_EHO model ($\text{LeuB}^{EX\_EHO}_{seq-only}$).

The Michaelis-Menten constants ($K_M$) for the substrate isopropylmalate (IPM) with $\text{LeuB}^{LG}_{joint}$ and $\text{LeuB}^{EX\_EHO}_{joint}$ are similar to those measured for contemporary LeuB enzymes (Table 2).

In contrast, the $K_M$ (IPM) for $\text{LeuB}^{EX\_EHO}_{seq-only}$ is >4-fold higher, showing its poorer affinity for this substrate (Table 2). Interestingly, $\text{LeuB}^{EX\_EHO}_{seq-only}$ exhibits a >2-fold higher turnover rate ($k_{cat}$) compared with the other two enzymes. Although this enzyme may have a high turnover rate, its high $K_M$ for IPM suggests that the substrate would have to be present at a very high concentration inside the cell for binding to occur.

The thermoactivity profiles of the three enzymes reveal that they are all highly thermophilic with $T_{opt}$ values > 75°C (Table 2 and Figure 4a).

In accordance with their high $T_{opt}$ values, both $\text{LeuB}^{LG}_{joint}$ and $\text{LeuB}^{EX\_EHO}_{joint}$ are very kinetically stable (as evidenced by their high values for $\Delta G^{\ddagger}_{N-U}$, which indicates the conformational stability of the molecule between the Native (folded) and the Unfolded states). However, $\text{LeuB}^{LG}_{joint}$ is much

| Enzyme | $K_M^{(IPM)}$ (mM) | $K_M^{(NAD)}$ (mM) | $k_{cat}$ ($s^{-1}$) | $T_{opt}$ ($^\circ C$) | $\Delta G^{\ddagger}_{N-U}$ ($kJmol^{-1}$) |
|---|---|---|---|---|---|
| BPSYC | 0.2 | 0.6 | 6.5 | 47 | 94.9 |
| BSUB | 0.7 | 8.1 | 48.7 | 53 | 95.9 |
| BCVX | 1.1 | 0.8 | 53.8 | 69 | 100.7 |
| ANC1 | 1.3 | 0.5 | 141.8 | 73 | 100.9 |
| ANC2 | 1.0 | 0.9 | 41.7 | 49 | 91.1 |
| ANC3 | 2.7 | 1.0 | 102.3 | 60 | 95.6 |
| ANC4 | 1.7 | 1.0 | 362.2 | 70 | 110.8 |
| **Joint Tree + LG** | **1.5** | **3.6** | **161.9** | **85** | **114.4** |
| **Joint Tree + EX_EHO** | **1.6** | **6.5** | **181.2** | **85** | **110.9** |
| **Sequence-only Tree + EX_EHO** | **6.8** | **5.5** | **441.2** | **78** | **91.4** |

Table 2: **Kinetic constants, thermoactivity and biophysical parameters for the ancestral LeuB enzyme from the Firmicutes ancestor.** Values obtained in this study for the ancestor of the Firmicutes (bold characters) were inferred using either the LeuB sequence tree or the LeuB reconciled tree and either with the site-homogeneous LG model or with the site-heterogeneous EX_EHO model. Data for contemporary (first three lines) and other ancestral LeuBs for Bacillus (ANC1-4) characterized in Hobbs et al. (2012) are shown for comparison

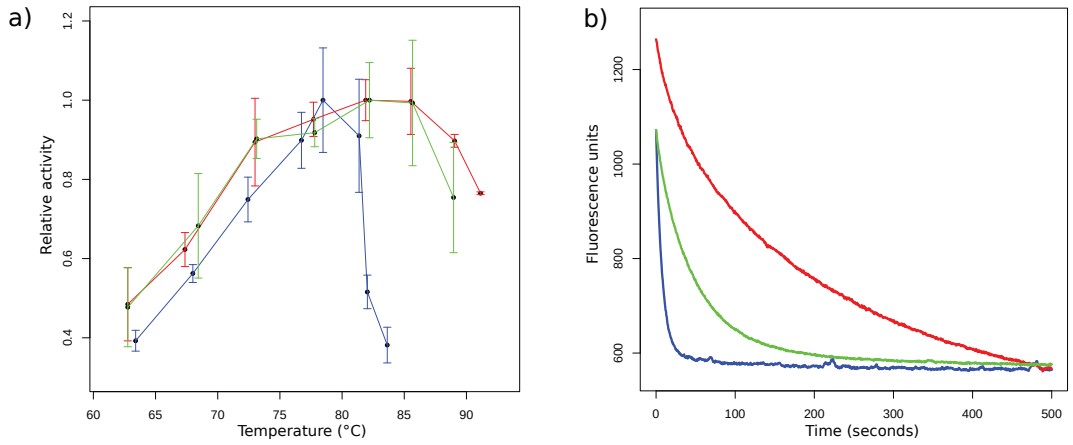



Figure 4: **Resurrection of ancestral LeuBs and impact of the substitution model and of the phylogenetic tree on biological interpretations**. a) Thermoactivity profiles for ancestral LeuB enzymes. Blue curve: sequence-only tree + EX_EHO. Red curve: joint tree + LG. Green curve: joint tree + EX_EHO. b) Unfolding rates of ancestral LeuB enzymes. Unfolding rates are shown in 8 M urea as a decrease in intrinsic protein fluorescence with time. Colors are the same as in a).

slower to unfold with respect to LeuB$^{EX_EHO}_{joint}$ (Figure 4b) and therefore is very kinetically stable. In contrast, LeuB$^{EX_EHO}_{seq-only}$ is thermophilic but kinetically unstable; its $\Delta G^{\ddagger}_{N-U}$ value is lower than that of contemporary and ancestral psychrophilic and mesophilic LeuB enzymes (Table 1). This suggests that, whilst LeuB$^{EX_EHO}_{seq-only}$ is adapted to function at high temperatures, it would unfold rapidly in a thermophilic environment. This instability, combined with the impaired $K_M$ for IPM, suggests that

this enzyme is not biologically realistic and implies that its inferred sequence may contain errors.

## Discussion

*In silico* ASR followed by experimental resurrection is a powerful approach to rewind evolutionary time and have access to the sequence and activity of ancestral proteins. A flourishing number of studies used this approach during the last decade. However, they were based on the use of simplistic substitution models and ML or Bayesian sequence-only trees (Harms and Thornton, 2010). The results presented here support the conclusion that the ASR and resurrection approach would benefit from recent methodological improvements in terms of substitution models and tree reconstruction.

We have shown that, depending on the model choice, the ML ancestral sequence and therefore the biological conclusions regarding its ancient properties, may vary greatly. We have demonstrated here that site-heterogeneous models systematically infer more accurate ancestral sequences. These results are in line with our previous study which showed that the use of time-heterogeneous models improved ASR accuracy when the evolutionary process varied between lineages (Groussin et al., 2013). On biological data, we have also shown that the use of more complex models can lead to changes in the ML ancestral state, even for residues that could have been considered as reconstructed without ambiguity with a homogeneous model. Furthermore, the average number of amino acid differences between the site-homogeneous LG model and site-heterogeneous models ranges from 7 to 20 on these data. This quantity is far from being negligible, especially as it is well known that a single or few mutations can lead to drastic changes in the structural stability or functional properties of a protein (Ortlund et al., 2007; Hobbs et al., 2013). Besides, if these residues are involved in solvent interactions, it may radically change the 3D configuration of the protein and distort biological conclusions regarding the co-evolution between the primary and the tertiary structure.

It has been previously suggested that consideration should be made with regard to evolutionary models that aim at capturing a greater part of the biological complexity of sequence evolution when attempting to perform ASR (Pupko et al., 2007; Hanson-Smith et al., 2010), as well as using model selection tests to choose the best-fitting model (Chang et al., 2002). In this study, we have shown that objective model selection criteria such as AIC allow the selection of the best substitution model regarding ASR accuracy. It is worth noting that with these simulated datasets as well as other biological datasets, the site-heterogeneous models almost systematically outperform models assuming homogeneity (Lartillot and Philippe, 2004; Le et al., 2008b,a; Le and Gascuel, 2010) in terms of data fitting. With the LeuB data, we have shown that model choice has an impact on final biological conclusions regarding kinetic stability (Figure 4b), although it is difficult to anticipate the impact of these differences the fitness of the organism. It is thus strongly recommended to include such models in the set of models

tested for data-fitting when attempting ASR. The Bio++ libraries (Dutheil et al., 2006; Guéguen et al., 2013) contain a large set of homogeneous, site-heterogeneous and time-heterogeneous models that can be easily used to compute ML estimates of evolutionary parameters and to infer ancestral sequences.

Finally, our *in silico* investigations strongly suggest that the use of a joint tree has a strong impact on the inference of ancestral sequences. This is confirmed by our resurrection experiment, which clearly illustrates the need for reconciled gene trees that maximise the joint sequence-reconciliation likelihood, as the inference performed with the sequence-only tree resulted in a realistic ancestor. When the gene family under study has experienced a complex evolutionary history involving gene duplications, lateral transfers and losses (such as LeuB), the effect of using a reconciled tree on the accuracy of ASR is even stronger than the effect of using a complex evolutionary model. Numerous methods that implement models of duplication, transfer and loss of genes are now available to reconcile a sequence-only tree with a species tree (David and Alm, 2011; Doyon et al., 2011; Wu et al., 2013; Rasmussen and Kellis, 2012; Szöllősi et al., 2013a). Here, we have demonstrated that the resulting gene tree is far more accurate than the original sequence- only tree and allows us to infer more accurately the history of protein evolution.

This study is not aimed at questioning previous biological conclusions obtained with the use of time- and site-homogeneous models to perform ASR along the sequence-only gene tree. However, previous (Groussin et al., 2013) and present results strongly suggest that the use of time- or site-heterogeneous models along reconciled gene trees should be in common use when performing ASR. This approach will provide access to accurate ancestral protein functions and structures.

# References


Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* pages 267–281. Petrov BN, Csaki F, editors Budapest (Hungary).

Åkerborg O, Sennblad B, Arvestad L, and Lagergren J. 2009. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc Natl Acad Sci U S A* 106:5714–5719.

Benner SA, Caraco MD, Thomson JM, and Gaucher EA. 2002. Planetary Biology–Paleontological, Geological, and Molecular Histories of Life. *Science* 296:864–868.

Boussau B, Szöllősi GJ, Duret L, Gouy M, Tannier E, and Daubin V. 2013. Genome-scale coestimation of species and gene trees. *Genome Res* 23:323–330.

Chang B and Donoghue M. 2000. Recreating ancestral proteins. *Trends Ecol Evol* 15:109–114.

Chang B, Jönsson K, Kazmi M, Donoghue MJ, and Sakmar TP. 2002. Recreating a Functional Ancestral Archosaur Visual Pigment. *Mol Biol Evol* 19(9):1483–1489.

Cole MF and Gaucher EA. 2011. Utilizing natural diversity to evolve protein function: applications towards thermostability. *Curr Opin Chem Biol* 15:399–406.

Criscuolo A and Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol* 10:210.

David LA and Alm EJ. 2011. Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* 469:93–96.

Doyon JP, Ranwez V, Daubin V, and Berry V. 2011. Models, algorithms and programs for phylogeny reconciliation. *Brief Bioinform* 12:392–400.

Dutheil J and Boussau B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol Biol* 8:255.

Dutheil J, Gaillard S, Bazin E, Glémin S, Ranwez V, Galtier N, and Belkhir K. 2006. Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinformatics* 7:188.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.

Finnigan GC, Hanson-Smith V, Stevens TH, and Thornton JW. 2012. Evolution of increased complexity in a molecular machine. *Nature* 481:360–364.

Fitch WM. 1971. Toward defining course of evolution–minimum change for a specific tree topology. *Syst Zool* 20:406–416.

Gaucher EA, Govindarajan S, and Ganesh OK. 2008. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* 451:704–708.

Gaucher EA, Thomson JM, Burgan MF, and Benner SA. 2003. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature* 425:285–288.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al.. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5:R80.

Groussin M, Boussau B, and Gouy M. 2013. A Branch-Heterogeneous Model of Protein Evolution for Efficient Inference of Ancestral Sequences. *Syst Biol* 62:523–538.

Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, Bigot T, Fournier D, Pouyet F, Cahais V, Bernard A, Scornavacca C, Nabholz B, Haudry A, Dachary L, Galtier N, Belkhir K, and Dutheil JY. 2013. Bio++: Efficient Extensible Libraries and Tools for computational molecular evolution. *Mol Biol Evol* 30:1745–1750.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, and Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59:307–321.

Hanson-Smith V, Kolaczkowski B, and Thornton JW. 2010. Robustness of Ancestral Sequence Reconstruction to Phylogenetic Uncertainty. *Mol Biol Evol* 27:1988–1999.

Harms MJ and Thornton JW. 2010. Analyzing protein structure and function using ancestral gene reconstruction. *Curr Opin Struct Biol* 20:360–366.

Hobbs JK, Jiao W, Easter AD, Parker EJ, Schipper LA, and Arcus VL. 2013. The change in heat capacity for enzyme catalysis determines the temperature dependence of enzyme catalysed rates. *ACS Chem Biol* doi:10.1021/cb4005029.

Hobbs JK, Shepherd C, Saul DJ, Demetras NJ, Haaning S, Monk CR, Daniel RM, and Arcus VL. 2012. On the Origin and Evolution of Thermophily: Reconstruction of Functional Precambrian Enzymes from Ancestors of Bacillus. *Mol Biol Evol* 29:825–835.

Höhna S and Drummond AJ. 2012. Guided tree topology proposals for bayesian phylogenetic inference. *Syst Biol* 61:1–11.

Jermann TM, Opitz JG, Stackhouse J, and Benner SA. 1995. Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature* 374:57–59.

Jones DT, Taylor WR, and Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8:275–282.

Katoh K and Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* 30:772–780.

Kodra JT, Skovgaard M, Madsen D, and Liberles DA. 2007. Linking sequence to function in drug design with ancestral sequence reconstruction. In *Ancestral Sequence Reconstruction* pages 34–39. Oxford University Press.

Konno A, Kitagawa A, Watanabe M, Ogawa T, and Shirai T. 2011. Tracing Protein Evolution through Ancestral Structures of Fish Galectin. *Structure* 19:711–721.

Koshi J and Goldstein R. 1996. Probabilistic reconstruction of ancestral protein sequences. *Journal of Molecular Evolution* 42:313–320.

Koshi JM and Goldstein RA. 1998. Models of natural mutations including site heterogeneity. *Proteins* 32:289–295.

Lartillot N, Lepage T, and Blanquart S. 2009. PhyloBayes 3. A Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25:2286–2288.

Lartillot N and Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol* 21:1095–2004.

Le SQ and Gascuel O. 2008. An Improved General Amino Acid Replacement Matrix. *Mol Biol Evol* 25:1307–1320.

Le SQ and Gascuel O. 2010. Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. *Syst Biol* 59:277–287.

Le SQ, Gascuel O, and Lartillot N. 2008a. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24:2317–2323.

Le SQ, Lartillot N, and Gascuel O. 2008b. Phylogenetic mixture models for proteins. *Phil. Trans. R. Soc. Lond. B* 363:3965–3976.

Löytynoja A and Goldman N. 2008. Phylogeny-Aware Gap Placement Prevents Errors in Sequence Alignment and Evolutionary Analysis. *Science* 320:1632–1635.

Malcolm B, Wilson K, Matthews B, Kirsch J, and Wilson A. 1990. Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. *Nature* 345:86–89.

Mirceta S, Signore A, Burns J, Cossins A, Campbell K, and Berenbrink M. 2013. Evolution of Mammalian Diving Capacity Traced by Myoglobin Net Surface Charge. *Science* 340(6138).

Miyata T, Miyazawa S, and Yasunaga T. 1979. Two types of amino acid substitutions in protein evolution. *J Mol Evol* 12:219–236.

Ortlund EA, Bridgham JT, Redinbo MR, and Thornton JW. 2007. Crystal structure of an ancient protein: evolution by conformational epistasis. *Science* 317:1544–1548.

Pauling L and Zuckerkandl E. 1963. Chemical Paleogenetics: Molecular "Restoration Studies" of Extinct Forms of Life. *Acta Chem Scand* 17:S9–S16.

Penel S, Arigon AM, Dufayard JF, Sertier AS, Daubin V, Duret L, Gouy M, and Perrière G. 2009. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics* 10:S3.

Pupko T, Doron-Faigenboim A, Liberles DA, and Cannarozzi GM. 2007. Probabilistic models and their impact on the accuracy of reconstructed ancestral protein sequences. In *Ancestral Sequence Reconstruction* pages 43–57. Oxford University Press.

Pupko T, Pe'er I, Shamir R, and Graur D. 2000. A Fast Algorithm for Joint Reconstruction of Ancestral Amino Acid Sequences. *Mol Biol Evol* 17:890–896.

Rasmussen MD and Kellis M. 2012. Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Res* 22:755–765.

Schneider R, de Daruvar A, and Sander C. 1997. The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res* 25:226–230.

Schroeder M, Culhane A, Quackenbush J, and Haibe-Kains B. 2011. Survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics* 27:3206–3208.

Stackhouse J, Presnell S, McGeehan G, Nambiar K, and Benner S. 1990. The ribonuclease from an extinct bovid ruminant. *FEBS Lett* 262:104–106.

Szöllősi GJ, Boussau B, Abby SS, Tannier E, and Daubin V. 2012. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc. Natl. Acad. Sci. U.S.A.* 109:17513–17518.

Szöllősi GJ, Rosikiewicz W, Boussau B, Tannier E, and Daubin V. 2013a. Efficient Exploration of the Space of Reconciled Gene Trees. *Syst Biol*.

Szöllősi GJ, Tannier E, Lartillot N, and Daubin V. 2013b. Lateral Gene Transfer from the Dead. *Syst Biol* 62:386–397.

Team RC. 2013. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*.

Thorne JL, Kishino H, and Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol* 15:1647–1657.

Voordeckers K, Brown CA, Vanneste K, van der Zande E, Voet A, Maere S, and Verstrepen KJ. 2012. Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. *PLoS Biol*. 10(12):e1001446.

Whelan S and Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* 18:691–699.

Williams PD, Pollock DD, Blackburne BP, and Goldstein RA. 2006. Assessing the accuracy of ancestral protein reconstruction methods. *Plos Comput Biol* 2:e69.

Wu YC, Rasmussen MD, Bansal MS, and Kellis M. 2013. TreeFix: Statistically Informed Gene Tree Error Correction Using Species Trees. *Syst Biol* 62:110–120.

Yang Z, Kumar S, and Nei M. 1995. A New Method of Inference of Ancestral Nucleotide and Amino Acid Sequences. *Genetics* 141:1641–1650.

# Supplementary material for:

# Biologically motivated models strongly improve the functionality of resurrected proteins

Mathieu Groussin[1], Joanne K Hobbs[2], Gergely J Szöllősi[1,3],

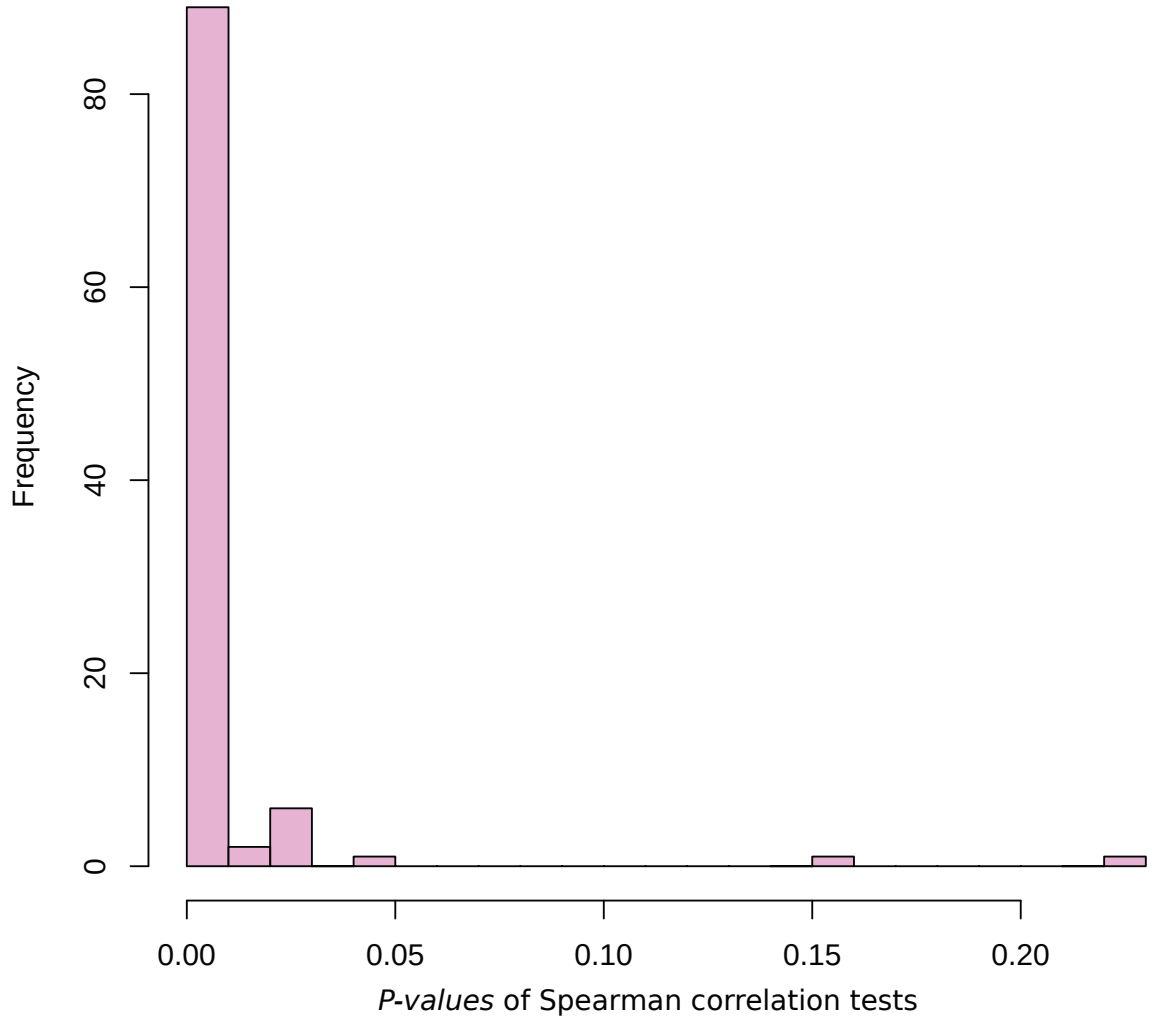
Simonetta Gribaldo[4], Vickery L. Arcus[2], and Manolo Gouy[1]

1 : *Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, Université Lyon 1, CNRS, UMR5558, Villeurbanne, France*

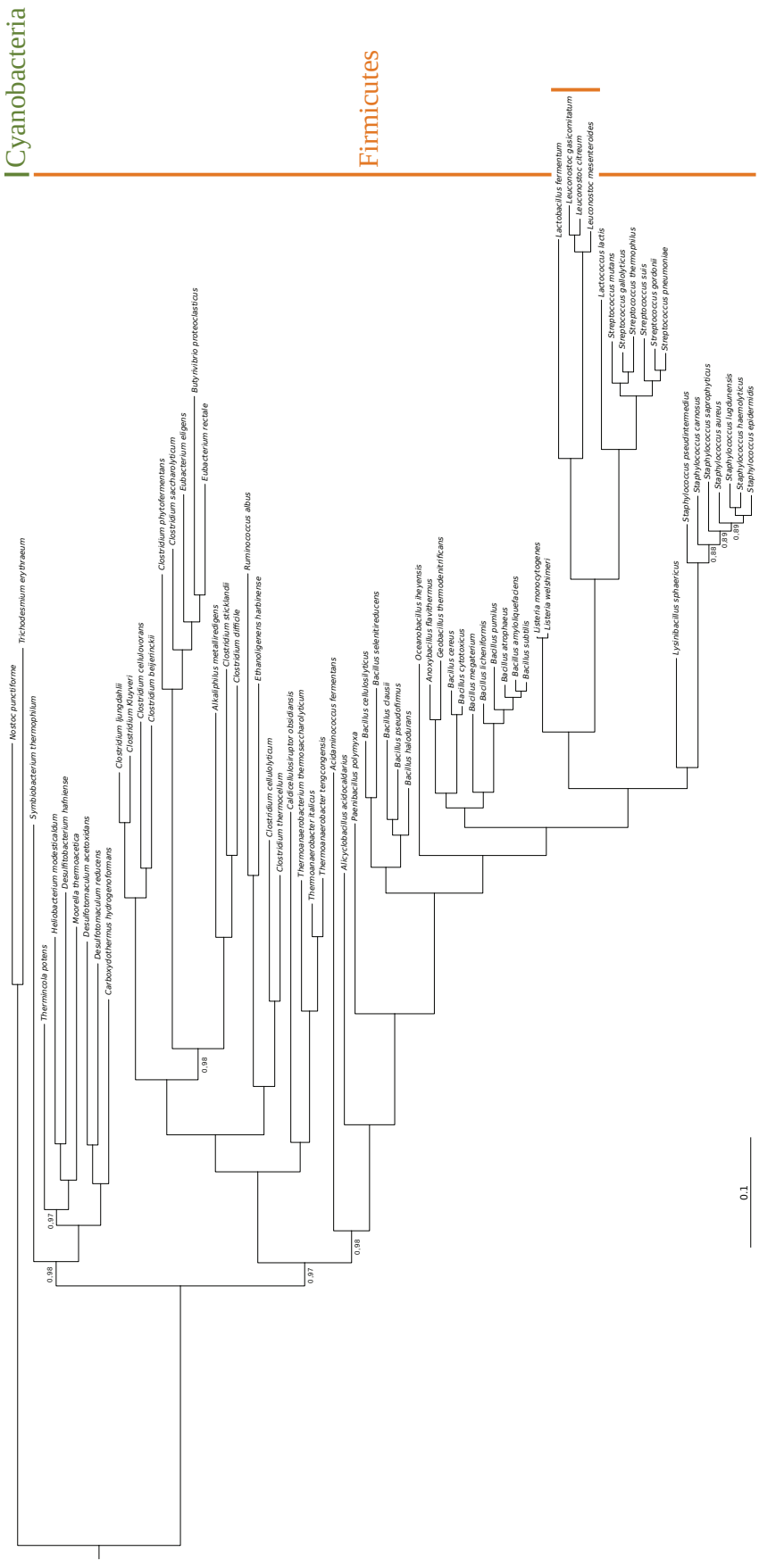
2 : *Department of Biological Sciences, University of Waikato, Hamilton, New Zealand*

3 : *ELTE-MTA "Lendület" Biophysics Research Group, Pázmány P. stny. 1A., H-1117 Budapest, Hungary*

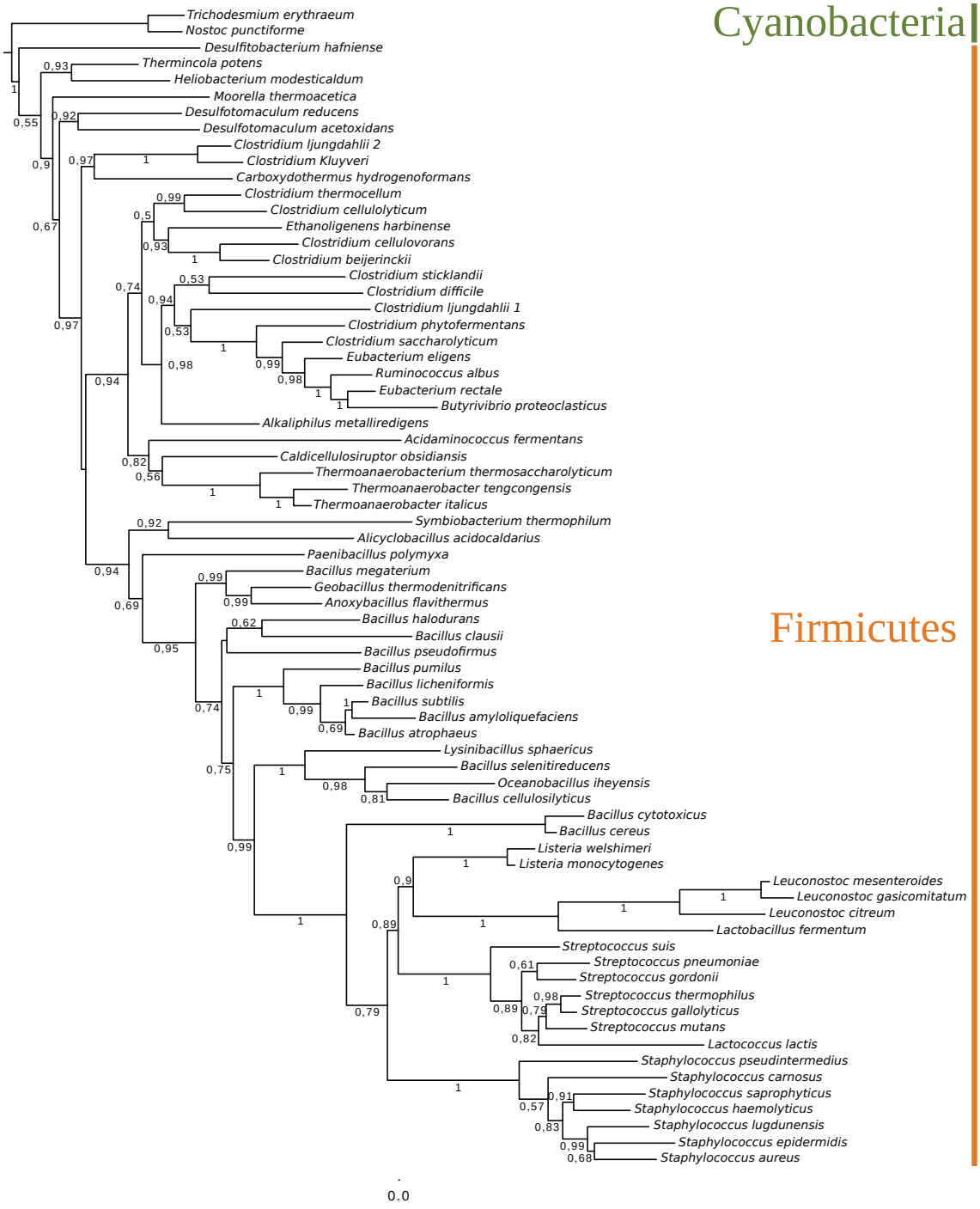
4 : *Institut Pasteur, Département de Microbiologie, Unité de Biologie Moléculaire du Gène chez les Extrêmophiles, 25-28 rue du Dr Roux, 75724 Paris cedex 15, France*



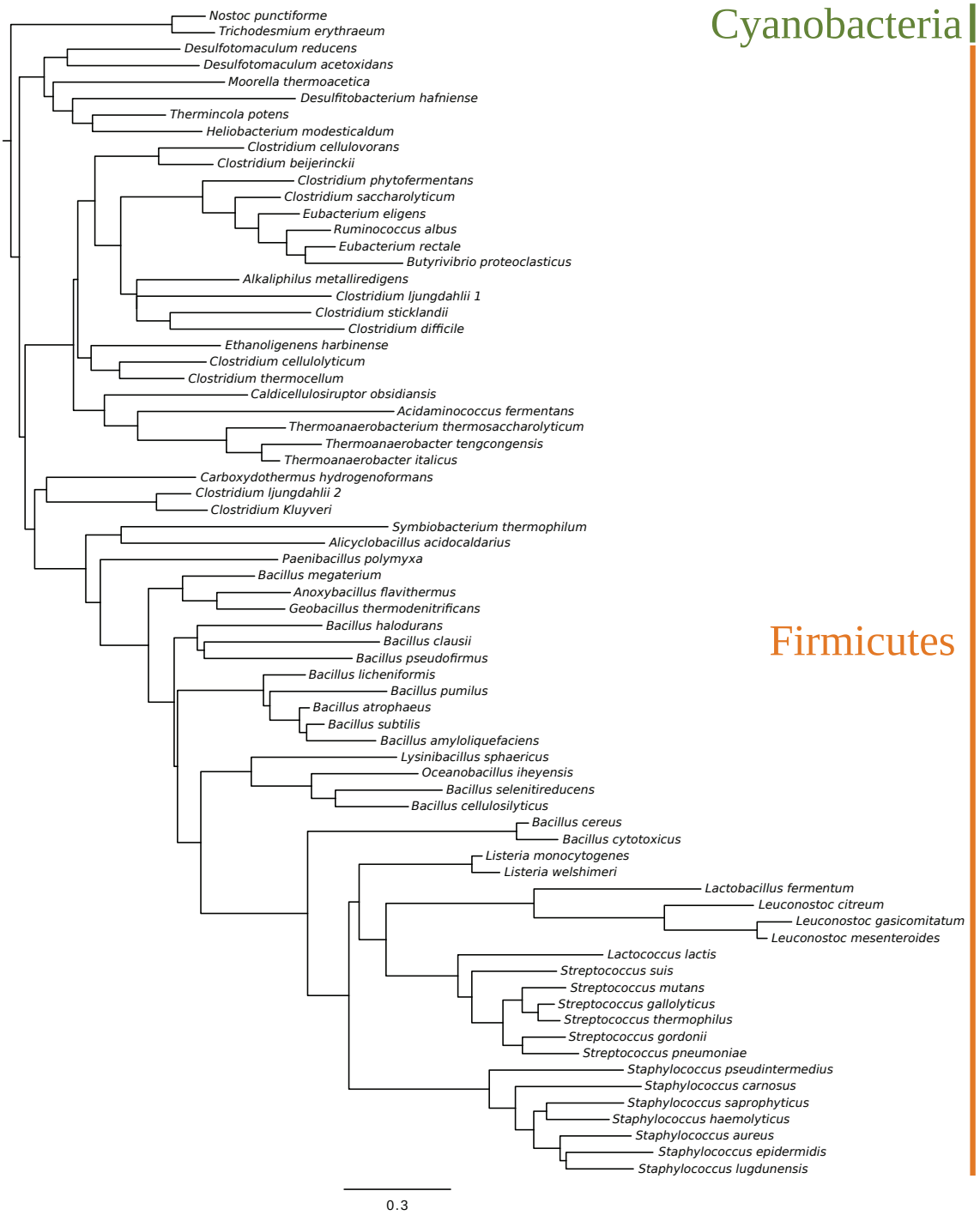

Supplementary Figure 1: **Correlation between the model fit to the data and the accuracy of ASR**. Several site-homogeneous and site-heterogeneous models were compared. For each simulated dataset, a Spearman correlation test was performed between the AIC value of the model and the average ASR accuracy (raw distance, see methods). This plot represents the distribution of the 100 *p-values*. The Fisher's combined probability test was then used to combine the results from the 100 independent Spearman correlation tests to test for a significant correlation.



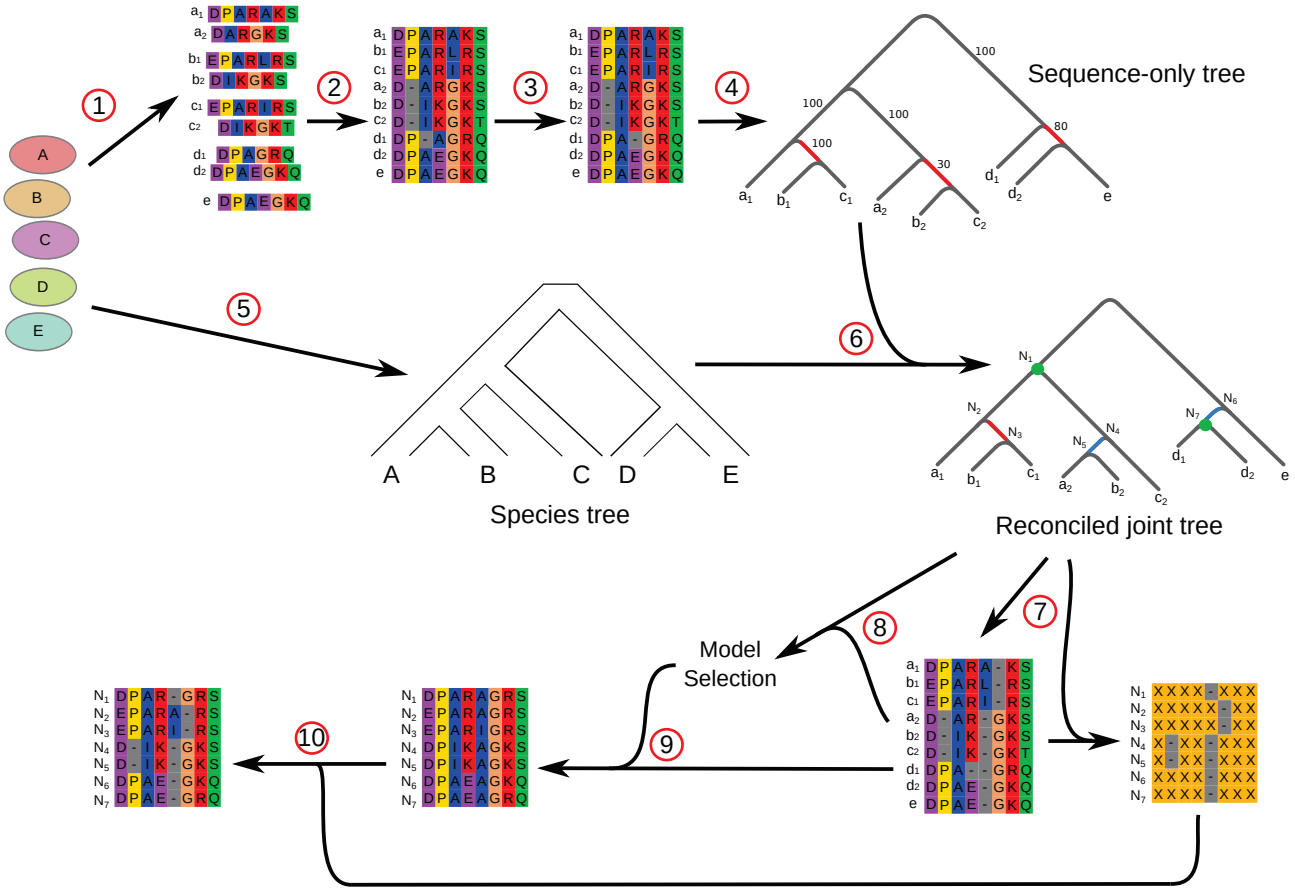

Supplementary Figure 2: **Species tree of Firmicutes**. This consensus posterior tree was reconstructed with the CAT model in Phylobayes. All branches with no support information have a PP of 1.




Supplementary Figure 3: **Sequence-only phylogenetic tree of LeuB sequences**. This tree represents the consensus posterior tree of LeuB sequences reconstructed with PhyloBayes, using the LG+Γ(4) model.




Supplementary Figure 4: **Reconciled joint phylogenetic tree of LeuB sequences**. This ML reconciled joint tree was reconstructed with ALE and represents the tree that maximises the joint sequence-reconciliation likelihood.




Supplementary Figure 5: **Standard protocol for ASR**. Here, a standard protocol for ASR is proposed, which accounts for the use of complex evolutionary models and gene tree/species tree reconciliation. The different steps are numbered in the same way as in the Online Material section. The species tree, sequence-only tree and reconciled tree are the same as in Figure 1.

## 4.2 Résurrections de protéines et adaptations structurales à l'halophilie.

### 4.2.1 Introduction

La façon dont les processus sélectifs ont façonnés la relation structure-fonction des protéines au cours du temps est encore peu comprise. Une des façons d'aborder cette problématique est d'utiliser la reconstruction et la résurrection de protéines ancestrales afin de déterminer leurs fonctions ancestrales et de les mettre en relation avec leur structure. Le manuscrit qui suit présente un tel projet, avec un intérêt tout particulier à l'adaptation aux conditions hypersalines des archées halophiles extrêmes. L'objectif est de décrire le chemin substitutionnel à travers lequel sont passées les ancêtres des Malates déhydrogénases (MalDH) des archées halophiles actuelles afin de déterminer l'influence de ces substitutions sur la fonction et le paysage conformationnel de cette enzyme soumise à des conditions environnementales extrêmes. Ce manuscrit fait écho au précédent, dans le sens où des modèles hétérogènes en sites et entre lignées ont été utilisés pour réaliser les inférences de séquences ancestrales le long d'un arbre réconcilié.

J'attire ici l'attention sur le fait que le manuscrit qui suit est dans un état préliminaire. L'approche *in silico* est, en revanche, aboutie et pleinement décrite. Des résultats d'expériences préliminaires menées sur un des ancêtres de la MalDH sont présentés. Dans le manuscrit final, une description complète et fine réalisée sur l'ensemble des protéines ancestrales ciblées le long de l'arbre des Haloarchaea sera présentée. Par conséquent, de nombreuses perspectives sont actuellement développées dans la discussion du manuscrit, permettant de donner une vision du projet général de résurrection envisagé.

### 4.2.2 Manuscrit

# Resurrection of halophilic proteins provides insights into the evolution of protein structure and function.

Mathieu Groussin[1,*], Samuel Blanquart[2,*], Gergely J Szöllősi[2,3], Manolo Gouy[1], and Dominique Madern[4]

*: These authors equally contributed to this work.

1 : *Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, Université Lyon 1, CNRS, UMR5558, Villeurbanne, France* 2 : *Inria Lille Nord Europe, LIFL UMR 8022 (CNRS Université de Lille 1), Villeneuve d'Ascq, France*

3 : *ELTE-MTA "Lendület" Biophysics Research Group 1117 Bp., Pázmány P. stny. 1A., Budapest, Hungary*

4 : *Dynamop Group, Institut de Biologie Structurale J.-P. Ebel, CEA CNRS UJF, UMR 5075, Grenoble cedex 01, France*

# Abstract

Extreme halophiles are organims adapted to hypersaline environments. This high salt concentration is usually mandatory for halophiles to grow and to have a proper metabolism. Halophilic proteins display particular adaptions at the amino acid level to maintain both solubility and functionality through the stabilisation of the native three-dimensional state. Among these adaptations, one can mention the increase of acidic residues at the surface of cytoplasmic proteins, as well as the presence of ion-binding sites, which are of great importance to stabilize the protein structure. However, how amino acid substitutions at the sequence level during evolution affect the conformational landscape and determine the evolvability of proteins is a major issue in evolutionary biology for which little is known. In this study, we propose to use the reconstruction and resurrection of ancestral MalDH enzymes in halophilic archaea to gain insight into the evolution of structure and function during the adaptive path to different halophilic conditions. MalDH is an adequate protein model because, despite the high identity between contemporary halophilic enzymes, different stability and functional properties are observed among Haloarchaea, increasing the possibility to link a particular substitution to a phenotypic change. To perform accurate ancestral sequence reconstruction, several substitution models have been compared in both Maximum Likelihood and Bayesian contexts, including biologically realistic models allowing the evolutionary process to vary among sites and in time. Furthermore, as several duplication, transfer and loss (DTL) events affected the MalDH history in Haloarchaea, we used a probabilistic model of gene tree/species tree reconciliation accounting for both sequence uncertainty and DTL events to infer the phylogenetic tree along which ancestral sequences were reconstructed. During a preliminary experiments, we resurrected the MalDH protein of the common ancestor of *Haloarcula marismortui* and *Haloferax volcanii*. We show that the protein folds properly and that it exhibits very different mechanisms of stability regulation with respect to protein-solvent interactions. These preliminary findings open the way to a full description of the influence of amino acid substitution on the evolution of catalysis and structure dynamics in relation with the adaptation to high salt concentrations.

# Introduction

The relationship between protein structure and protein function has been the matter of much investigations during the last decades (Lee et al., 2007). Understanding this relationship is key to fully appreciate evolvability properties of proteins and how they acquire new structures and functions during evolution (Tokuriki and Tawfik, 2009). Both structure and function are determined by the amino acids sequence and the environmental conditions in which the protein resides in the cell. It may be thought that a given protein has only one well-defined structure directly linked to a strong functional specificity. However, it is now acknowledged that a given protein exists as a population of conformers, defined as alternative substructures having slightly different free energies and being statistically represented in a specific proportion (Dill and Chan, 1997; Frauenfelder et al., 2009). In vivo, the conformational landcape of a given protein is partly determined by intra-monomeric (and possibly inter-monomeric, in the case of multimeric protein assemblages) interactions between residues through hydrophobic or ionic interactions, hydrogen bonds or Van der Waals forces. But protein folding is also influenced by solvent interactions, such that proteins should be rather seen as dynamical protein-solvent complexes (Zaccai, 2004). Therefore, the global stability and the equilibrium between conformational substates of a protein in natural conditions are a tripartite entity that depends on both stabilizing and destabilizing interactions and solvent interactions (Ebel et al., 1999; Jaenicke, 2000)

To investigate the complex relationship between all these interactions, halophilic proteins have been proved to be very informative entities (Irimia et al., 2003). Halophiles, which have been reported in Bacteria and Archaea, are organisms adapted to hypersaline environments and can grow in very high salt concentrations (up to 4 M NaCl or even higher) that would normally destabilize and make unsoluble any regular non-halophilic protein (Madern and Zaccaï, 1997). Two strategies may be employed to adapt to such extreme conditions. The first strategy is to synthesize and accumulate osmoprotectants (ca. polyols, betaine, etc) to increase the osmolarity of the cell. The second strategy is to accumulate potassium ($K^+$) within the cytoplasm, in turn requiring molecular adaptation at the protein level to face the high concentrations of cations ($K^+$) and anions ($Cl^-$) (Madern et al., 2000). In line with this, it has been shown that proteins purified from these organisms can be defined as halophilic proteins specifically adapted to hypersaline environments, because they are only stable, soluble and active at mildly- or highly-elevated salt concentrations (Irimia et al., 2003). A substantial part of the molecular adaptation to high salt concentrations concerns the enrichment in acidic residues at the surface of the protein to interact with water molecules and $K^+$ ions (Richard et al., 2000) so that the protein remains soluble (Baliga et al., 2004). Another molecular adaptation concerns a limited number of sites that strongly bind solvent anions to highly increase protein stability (Madern et al., 2007).

One of the most documented halophilic protein in the litterature is the L-Malate dehydrogenase

(MalDH). MalDH is an enzyme involved in the citric acid (Krebs) cycle that reversibly catalyses the conversion of malate into oxaloacetate with a NAD(P)H-dependent oxidation of malate. MalDH is part of a larger family of dehydrogenases comprising NADH-dependent L-Lactate dehydrogenases and alcohol dehydrogenases (Birktoft et al., 1982). MalDH-encoding genes have been universally reported in Bacteria, Archaea and Eukaryotes. MalDH is a multimeric enzyme with subunit molecular masses of 3040 KDa that can be found in dimeric or tetrameric states (Sundaram et al., 1980; Richard et al., 2000), with both states being able to perform enzymatic activity (Madern et al., 2001). The crystal structure of the *Haloarcula marismortui* MalDH has been determined at 3.2 Å resolution (Richard et al., 2000; Irimia et al., 2003). The tetrameric structure shows that $Cl^-$ anions form salt bridges with residues of MalDH at the dimer-dimer interface (Dym et al., 1995), enhancing protein stability in the high salt concentrations in which *H. marismortui* lives. It strongly emphazises the role of protein-solvent interactions at the interface between monomers in the adaptation to halophilic conditions.

Protein structure and function are primarily determined at the level of the amino-acid sequence, upon which the substitution process acts ((Liu and Bahar, 2012)). Extant sequences record this substitution history, which can be recovered with probabilistic substitution models, leading to the possibility to infer what were their ancestral sequences. Ancestral protein sequence resurrection is a powerful approach to have access to ancient protein properties such as structure and function and how their relationship with the primary sequence evolved in time through successive substitutions (Harms and Thornton, 2010). The *in-vitro* and/or *in-vivo* resurrection of ancestral proteins depends on the *in-silico* inference of ancestral proteins from sequences present in extant organisms. With an alignment of homologous sequences, their corresponding phylogenetic tree and using a substitution model providing substitution probabilities between the different amino acids, ancestral sequences can be computed at each internal node of the tree. Subsequently, these sequences may be synthesized and expressed in a cell, providing access to extinct molecules. Since the beginning of the 1990s Malcolm et al. (1990) and Stackhouse et al. (1990), many studies used ancestral sequence reconstruction (ASR) to investigate ancient adaptations to temperature among Bacteria (Gaucher et al., 2003; Hobbs et al., 2012), ancestral ecological adaptations to light (Chang et al., 2002), the evolution of diving capacity in Mammals (Mirceta et al., 2013), the influence of gene duplication on functional divergence (Voordeckers et al., 2012), the evolution of molecular complexes (Finnigan et al., 2012) or drug design (Kodra et al., 2007). It has been recently shown that the realism of the substitution model is correlated with ASR accuracy (Groussin et al., 2013a,c). For instance, substitution models that consider the substitution process to be homogeneous along the sequence or in time were proved to estimate less accurate ancestral sequences than models that allow the process to vary among sites (Lartillot and Philippe, 2004; Le et al., 2008b; Groussin et al., 2013b) or between lineages (Foster, 2004; Blanquart and Lartillot, 2006; Groussin et al., 2013a) (henceforth named site-heterogeneous and time-heterogeneous models, respectively). Furthermore, Groussin et al. (2013c) showed that the phylogenetic tree along which ancestral sequences are computed has a strong impact

on ASR accuracy. Indeed, beyond the evolution at the substitution level, a protein also evolves in terms of duplications, transfers and losses (DTL) through time. These events generally lead to phylogenetic trees reconstructed from the sequences (henceforth named *sequence-only* trees) by classic Maximum Likelihood (ML) or Bayesian Inference (BI) methods that are inconsistent with the topology of the species tree. Topological inconsistency may also arise from the lack of phylogenetic signal present in the gene alignment, preventing to obtain well-supported divergences. Several methods were proposed to reconcile the sequence tree with the species tree (Åkerborg et al., 2009; Rasmussen and Kellis, 2012; Wu et al., 2013; Boussau et al., 2013; Szöllősi et al., 2013a) and to consider the sequence information to detect unsupported branches and retain bona fide phylogenetic discord produced by genome evolutionary proceses. These *joint* reconciled tree topologies have been proved to be far more accurate than sequence-only tree topologies (Szöllősi et al., 2013a). The reconciliation process is of particular interest in our case, as it was shown that numerous horizontal gene transfers occurred during the evolution of Haloarchaea (Williams et al., 2012).

In this study, we propose to use ASR and resurrection of extant haloarchaeal MalDH sequences to gain insights into the specific features that contributed to protein stability and solubility during adaptation to various halophilic conditions. We used a large taxonomic sampling of haloarchaeal and outgroup species. We controlled for the influence of the statistical framework in which ASR was performed (ML vs. Bayesian), as well as for the influence of the substitution model on the estimation of ancestral MalDH sequences. A species tree reconstruction was performed, yielding to a fully resolved Haloarchaea tree. The species and MalDH trees were subsequently used by a species tree/sequence tree reconciliation algorithm. MalDH ancestral sequences were then reconstructed along this reconciled joint tree, which is, to our knowledge, the first attempt in using joint trees to perform ASR. Preliminary biochemical experiments performed on a key ancestor show that the ancestral protein folds properly and has different stability properties with respect to extant MalDH of *H. marismortui* and *H. volcanii*, due to a different pattern of solvent interactions. Future experiments should allow us to more deeply comprehend how substitutions modified the interactions involved in the determination of the conformational landscape of MalDH during adaptation to halophilic environments.

# Material and Methods

## Selection of species

### Ingroup

As of April 2011, 12 haloarchaeal genome sequences are available in GenBank. To this set of species, we added haloarchaeal genomes downloaded from `http://www.bme.ucdavis.edu/facciotti/resources_`

`data/data/haloarchaeal-genomes/` to obtain a total of 55 haloarchaeal genomes. The malate dehydrogenase gene family was then reconstructed with Silix (Miele et al., 2011) (see below) and species lacking MalDH were removed from the dataset. Besides, several genomes contain nearly identical MalDH sequences. We randomly selected a MalDH sequence among these species to avoid taxonomic redundancy and controled for the influence of the random choice on the reconstruction of the species tree. The MalDH sequences for which a 3D structure is available were conserved. Finally, three species (*Halobacterium salinarum R1*, *Natronomonas pharaonis DSM 2160* and *Halalkalicoccus jeotgali B3 5435*) were discarded from the dataset due to difficulties to confidently decipher their phylogenetic position in the species tree. 28 haloarchaeal species remained in the final dataset. Only two species (*Haloterrigena thermotolerans* and *Natronolimnobius innermongolicus*) possess a duplicate gene of MalDH.

### Outgroup

A taxonomically-rich outgroup was considered during the reconstruction of the species tree and MalDH trees. We considered the archaeal species tree reconstructed in (Brochier-Armanet et al., 2011) to select 21 euryarcheal species outside of Haloarchaea. We retained the recently sequenced genomes of the two nanohaloarchaeal species (*Candidatus Nanosalinarum sp. J07AB56* and *Candidatus Nanosalina sp. J07AB43*), as it was suggested that they diverged just before the appearance of Haloarchaea (Narasingarao et al., 2012).

## Reconstruction of gene families

From the collection of 51 proteomes, homologous gene families have been obtained with an all-against-all Blast approach. The program Silix (Miele et al., 2011) was then used to cluster protein sequences into homologous families. Default parameters were considered for the clustering: sequences having more than 80% similarity and more than 30% coverage were clustered together into a homologous gene family. Uni-copy gene families having more than 80% of taxonomic coverage were conserved. These 240 gene families were further aligned with Prank (Löytynoja and Goldman, 2005, 2008), internally used in Guidance (Penn et al., 2010) to trim ambiguously aligned sites by taking into account the uncertainty of the guide tree during the alignment procedure. Ambiguously aligned sites were further trimmed with Gblocks (Castresana, 2000), with standard options and with gaps allowed. Finally, the bppSeqMan program belonging to the bppSuite of programs (Dutheil and Boussau, 2008) was used to eliminate sites containing more than 20% of gaps.

| Phylum | Species name |
|---|---|
| Haloarchaea | Halorhabdus utahensis DSM 12940 |
| | Halosimplex carlsbadense |
| | Haloarcula marismortui ATCC 43049 |
| | Halomicrobium mukohataei DSM 12286 |
| | Halorubrum hochstenium ATCC 700873 |
| | Halorubrum arcis |
| | Halorubrum californiensis |
| | Halorubrum aidingense |
| | Halorubrum lacusprofundi ATCC 49239 |
| | Haloferax volcanii DS2 |
| | Halogeometricum borinquense DSM 11551 |
| | Haloquadratum walsbyi DSM 16790 |
| | Halovivax asiaticus |
| | Natronococcus amylolyticus DSM 10524 |
| | Natronococcus jeotgali |
| | Natrialba aegyptia DSM 13077 |
| | Natrialba magadii ATCC 43099 |
| | Haloterrigena limicola JCM 13563 |
| | Haloterrigena thermotolerans |
| | Natrinema altunense |
| | Natronorubrum tibetense |
| | Natronorubrum bangense |
| | Natronorubrum sulfidifaciens JCM 14089 |
| | Haloterrigena turkmenica DSM 5511 |
| | Natronolimnobius innermongolicus |
| Nanohaloarchaea | Candidatus NanoSali Sp |
| | Candidatus NanoSalinarium Sp |
| Methanocellales | Methanocella paludicola SANAE |
| | Uncultured Methanogenic Archaeon RCI |
| Methanosarcinales | Methanosaeta thermophila PT |
| | Methanosaeta concilii GP6 |
| | Methanosarcina barkeri str. Fusaro |
| | Methanosarcina mazei Go1 |
| | Methanosarcina acetivorans C2A |
| | Methanococcoides burtonii DSM 6242 |
| | Methanohalophilus mahii DSM 5219 |
| | Methanohalobium evestigatum Z 7303 |
| | Methanosalsum zhilinae DSM 4017 |
| Methanomicrobiales | Methanocorpusculum labreanum Z |
| | Methanoplanus petrolearius DSM 11571 |
| | Methanoculleus marisnigri JR1 |
| | Methanospirillum hungatei JF-1 |
| | Candidatus Methanoregula boonei 6A8 |
| | Methanosphaerula palustris E1 9c |
| Archaeoglobales | Ferroglobus placidus DSM 10642 |
| | Archaeoglobus fulgidus DSM 4304 |
| | Archaeoglobus profundus DSM 5631 |
| | Archaeoglobus veneficus SNP6 |
| Thermoplasmatales | Thermoplasma volcanium GSS1 |
| | Thermoplasma acidophilum DSM 1728 |
| | Picrophilus torridus DSM 9790 |

Table 1: Species considered in the protein concatenate

## Reconstruction of the species tree

### Reconstruction of the species tree with ribosomal RNAs

### Phylogenomic reconstruction of the species tree

The species tree has been inferred with a concatenation approach. After the concatenation of the 240 gene families, the final alignment contains 50,135 amino-acid positions. Different strategies were employed to reconstruct the species tree. The LG substitution model (Le and Gascuel, 2008) and a $\Gamma$ distribution with 4 categories were used in PhyML (Guindon and Gascuel, 2003; Guindon et al., 2010) to reconstruct the tree with 100 bootstraps. The LG model assumes that the evolutionary process is constant between lineages and across sites. The CAT and CAT-GTR (Lartillot and Philippe, 2004) models implemented in Phylobayes (Lartillot et al., 2009) were also employed. CAT and CAT-GTR assume that the process is heterogeneous among sites and constant between lineages. We computed 1,000,000 cycle long MCMC chains, saving a sample each 10 cycles. The 1,000 first samples were discarded as burnin. Two independent chains were executed for each experiment, and the chain's convergence was assessed if tree bipartition differ by $PP < 0.1$. Finally, the COaLA substitution model was also used with bppML (Dutheil and Boussau, 2008). COaLA is site-homogeneous but implements a time-heterogeneous model that allows to model the variation of global compositions between lineages. As topology exploration is not feasible with time-heterogeneous models in bppML, COaLA was used to test alternative topologies regarding the position of the root of the Haloarchaea clade. To discriminate between alternative topologies, AU tests were performed with Consel (Shimodaira and Hasegawa, 2001).

To improve the resolution of the phylogeny of Haloarchaea, and especially the position of its root, an elimination of the fast-evolving positions was realised to reduce systematic errors (Philippe et al., 2005; Brinkmann et al., 2005). We used the site-specific posterior rates computed by PhyML with the use of the $\Gamma$ distribution to gradually remove fast-evolving sites (by fractions of 10% of sites).

## Reconstruction of the MalDH sequence-only tree

The MalDH gene family was reconstructed with Silix (Miele et al., 2011). The 51 MalDH sequences were then aligned with Muscle (Edgar, 2004), used in Guidance (Penn et al., 2010), which allows to trim ambiguously aligned sites owing to uncertainty in the guide tree topology. Remaining ambiguously aligned sites were removed with Gblocks (Castresana, 2000).

The Bayesian sequence-only tree was reconstructed using the CAT-GTR model implemented in PhyloBayes 3.0 (Lartillot et al., 2009). We ran 1,000,000 long MCMC chains, saving a sample each 10 cycles, and we discarded the 1,000 first samples as burnin. The consensus sequence-only tree was then

reconciled with the species tree.

## Species tree-Gene tree reconciliation

A pre-released version of the Amalgamated Likelihood Estimation (ALE) algorithm (ALE v0.1) (Szöllősi et al., 2013a) was used to compute the ML MalDH joint tree. ALE uses the model described in Szöllősi et al. (2013b) to search for the best scenario of duplications, transfers and losses of genes and efficiently explores the space of joint trees that maximise the joint likelihood between the sequence and reconciliation information. ALE requires a time calibrated species tree to compute ML estimates of transfer rates. Divergence times were estimated using PhyloBayes from the genomic concatenation having 50% of its fastest evolving sites filtered out (see above). We used the CAT substitution model, the Log Normal relaxed molecular clock, no calibration point, and a flat root prior defined by a 1 billion years expectation and standard deviation. The species tree was rooted with the Thermoplasmatales. The MCMC chains were elongated for 100,000 cycles, a sample was saved every 10 cycles, and 1,000 first samples were discarded as burnin.

ALE also uses a sample of gene trees to compute conditional clade probabilities (Höhna and Drummond, 2012), which can be used to approximately estimate the posterior probability of a gene tree that can be amalgamated from clades present in the sample. The sample of posterior trees computed by PhyloBayes on the MalDH alignment (see above) was provided to ALE.

## Ancestral Sequence Reconstruction of MalDH

The MalDH joint tree was used as a final guide tree to Prank, which is very sensitive to the choice of the guide tree (Löytynoja and Goldman, 2008). The Prank alignment and the joint tree were used to compute the ancestral sequences. Moreover, one of the key outputs of Prank is the inferred history of the insertion/deletion events leading to the extant gap pattern. This history is used to filter out from the computed ancestral sequences all ancestral sites that are inferred as gaps by Prank.

To reconstruct ancestral MalDH sequences in ML, the bppAncestor program (Dutheil and Boussau, 2008) was used with the marginal reconstruction approach (Yang et al., 1995). For a given node at a given site, bppAncestor makes use of the ML estimates of branch lengths and models parameters obtained with bppML to compute the posterior probabilities (PP) of each possible ancestral states. The state having the maximum posterior probability is inferred as being the ML ancestral state. Site- and time-homogeneous models (LG and LG+$F_{opt}$), site-heterogeneous (EX2, EX3, EHO, UL2, UL3 (Le et al., 2008b), EX_EHO (Le and Gascuel, 2010), C10 to C60 (Le et al., 2008a)) and time-heterogeneous models (COaLA with 1 or 2 parameters per branch (Groussin et al., 2013a)) were run with bppML to

retain the best-fitting model in terms of AIC (Akaike, 1973) and BIC (Schwarz, 1978) values. Ancestral gaps inferred by Prank were then incorporated in the final ancestral sequences.

Several models available in PhyloBayes were used to compute the ancestral sequences: LG, GTR, CAT and CAT-GTR. As the latter model generally provides the best fit to the data, its estimation provided us the ancestral sequences to be synthesized. MCMC chain's were run for 1,000,000 cycles, saving a sample each 10 cycles and discarding the first 1,000 sample as burnin. Two independent chains were checked for convergence using the "tracecomp" program provided in PhyloBayes. The ancestral sequences were computed from the posterior distributions using the "ancestral" program of the PhyloBayes suite. We used the mid point rooting option (-midpointrooting) to provide an estimate at the root node. As final step of the ancestral sequence inference, ancestral gaps inferred by prank are substituted to the inferred ancestral states, yielding the ancestral amino acid sequences to be synthesized.

# Results

## Species tree reconstruction

Figure 1 shows that the ML Haloarchaea phylogeny inferred with PhyML and the LG model is well resolved. Although a lot of short branches are present in Haloarchaea, the diversification pattern of this group is strongly supported by the bootstrap analysis. Nonetheless, uncertainty remains regarding the position of the root of Haloarchaea using the whole concatenate. With 100% of the sites, the bootstrap support for the diversification between clades A (*Halorhabdus utahensis, Halomicrobium mukohataei, Haloarcula marismortui* and *Halosimplex carlsbadense*) and B+C is weak (55%). The two Nanohaloarcheal species are placed at the base of Haloarchaea, as suggested by previous phylogenetic analyses of rRNAs (Narasingarao et al., 2012). The very long branches leading to both Haloarchaea and Nanohaloarchaea groups probably explain the weak resolution for the first divergence of Haloarchaea. It is worth noting that a well-supported species topology is necessary to perform species tree/gene reconciliation and to further infer ancestral MalDH sequences of Haloarchaea. Consequently, the reason for this uncertainty was investigated. To decrease systematic errors that make the site- and time-homogeneous LG model (Le and Gascuel, 2008) more prone to incorrectly discriminate vertical descent from convergence or reversion events due to multiple substitutions, we gradually removed fast-evolving sites from the concatenate (see Material and Methods) (Philippe et al., 2005; Brinkmann et al., 2005).

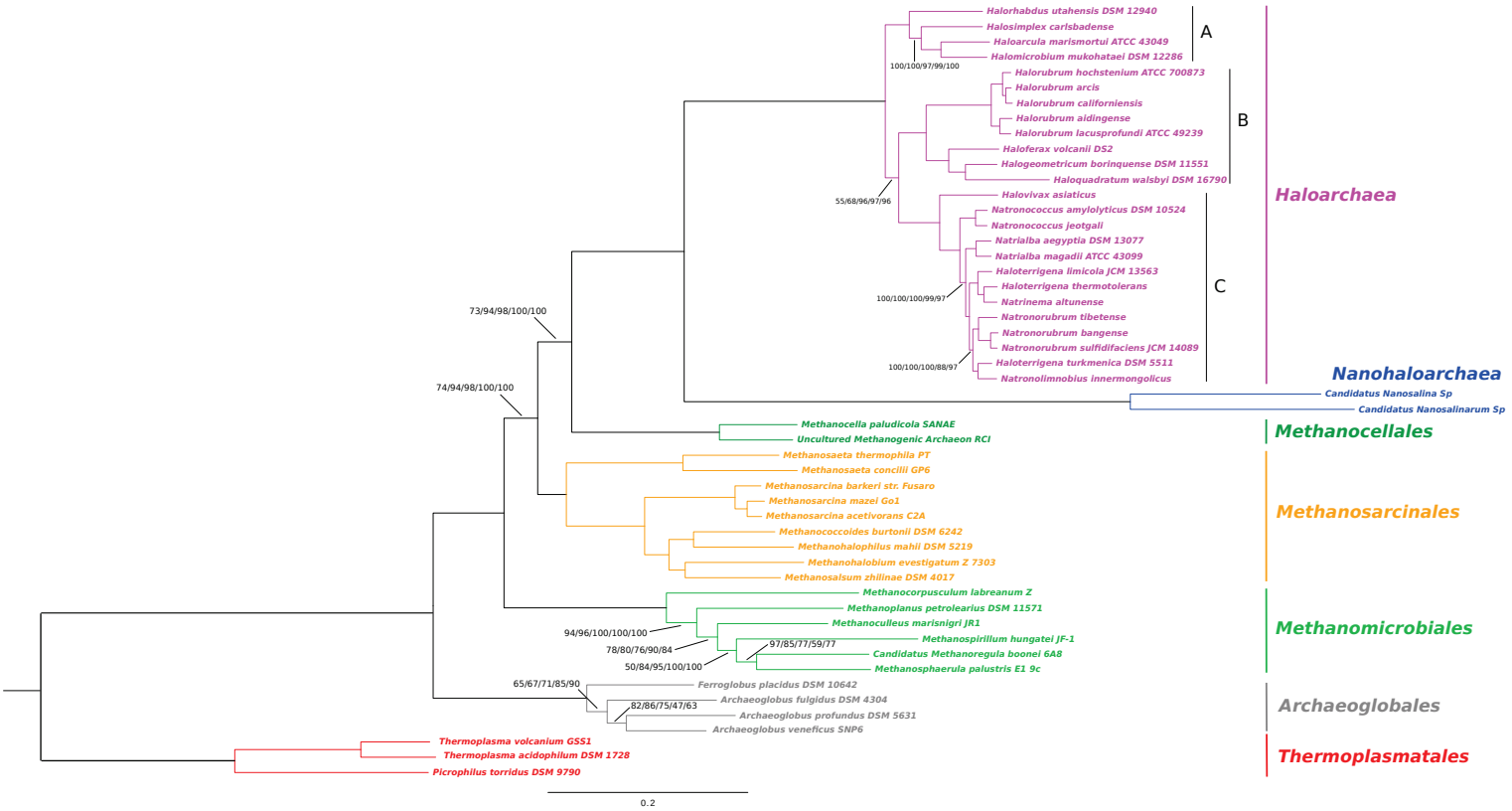



Figure 1: **Phylogenetic tree of the 51 archaeal species**. The tree was reconstructed with a ML approach using PhyML and the LG+Γ(4) model. The alignment contains 51 archaeal species and 50,135 sites. The different phyla are color-coded. The support for the tree was evaluated with a bootstrap approach, using 100 replicates. Branches with no support information have 100% bootstrap support. When bootstrap values are indicated, they correspond to the result obtained with an elimination of 0%/10%/20%/30%/40% of the fast-evolving sites originally present in the dataset. In each case, the ML topology was inferred to be same. The species was rooted following Brochier-Armanet et al. (2011).

Figure 2 shows that the elimination of 20% of fast-evolving sites results in a strong confidence for the position of the root of Haloarchaea, in agreement with the low-supported position found with the whole alignment. Previous phylogenetic analyses performed on Haloarchaea or Archaea at a larger scale did not allow to decipher the position of the haloarchaeal root (López-García et al., 2001; Brochier-Armanet et al., 2008, 2011). The present phylogenomic analysis yields a strongly-supported position that differs from the position proposed by Brochier-Armanet et al. (2011) where the first emerging clade is B, with strong support. However, Brochier-Armanet et al. (2011) considered less phylogenetic information to reconstruct their tree, with a limited taxonomic sampling for Haloarchaea, the absence of Nanohaloarchaea and a reduced number of genes and positions (57 ribosomal proteins were present in their final alignment, containing 5838 sites). Finally, their tree was obtained with a Dayhoff-6 amino acid recoding procedure, which is usually employed to reduce compositional biases (Hrdy et al., 2004; Rodriguez-Ezpeleta et al., 2007).

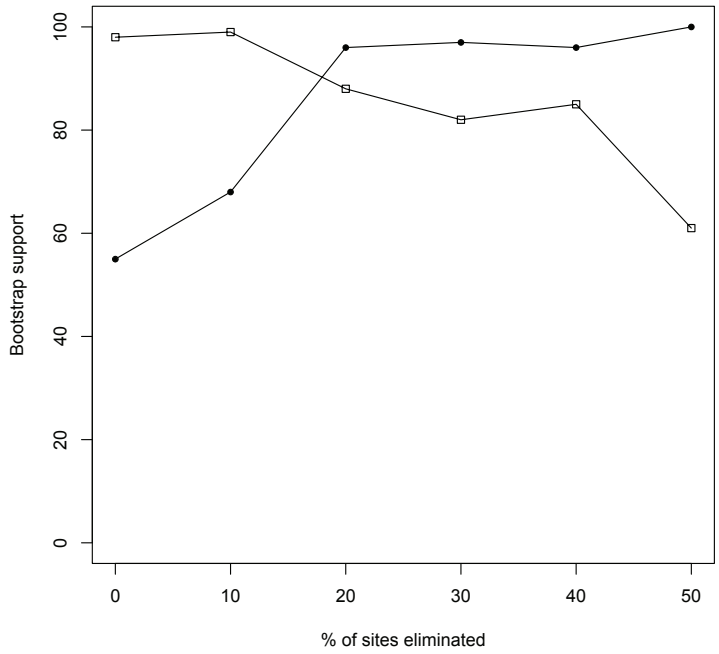



Figure 2: **Gradual removal of fast-evolving sites and impact on the position of the Haloarchaea root**. The curve with black dots refers to the support for the haloarchaeal root placing group A at the base of Haloarchaea (see Figure 1). The curve with open squares correspond to the support for the root with B emerging first, obtained with the Dayhoff-6 recoded dataset (see Materials and Methods section).

We also performed phylogenomic reconstructions with the same recoding procedure. Figure 2 shows that with 100% of sites, the root position of Brochier-Armanet et al. (2011) is strongly supported. However, the more fast-evolving sites are removed, the less this position is supported. Furthermore, we used the COaLA model (Groussin et al., 2013a) on the non-recoded alignment to compute the likelihood of the three alternative rearrangements of A, B and C around the root of Haloarchaea. COaLA is time-heterogeneous and is able to efficiently model the variation of composition between lineages by considering only a few parameters to optimize per branch. Table 2 shows that COaLA confirms the first results obtained with LG on the non-recoded datasets, underlining that composition biases are unlikely to explain the uncertainty for the root position.

| | AU values | | |
|---|---|---|---|
| % of sites eliminated | Root A | Root B | Root C |
| 0% | 0.728 | 0.295 | 0.005 |
| 10% | 0.809 | 0.198 | 0.001 |
| 20% | 0.961 | 0.040 | 2e-04 |
| 30% | 0.982 | 0.018 | 3e-04 |
| 40% | 0.985 | 0.015 | 1e-04 |
| 50% | 0.996 | 0.004 | 2e-04 |

Table 2: **Topology comparison for the inference of the root position of Haloarchaea with COaLA**. Root X correspond to a root where group X emerges first in Haloarchaea (see Figure 1). Values in the table correspond to the results of AU tests of topology comparisons.

Finally, the genomic concatenations were also analyzed under a Bayesian framework, but many MCMC chains failed to reach convergence, especially with the biggest datasets (not shown). Nonetheless, on the genomic concatenation having 50% of its fastest evolving sites filtered out, all the four MCMC chains converged with the site heterogeneous CAT model. A collection of 128180 trees was used to compute the posterior tree (see Materials & Methods). Phylogenetic relationships within Haloarchaea are all strongly supported with $PP = 1$ and are all congruent with the ML conclusions (Figure 3). The CAT model is known to be less sensitive to long branch attraction artifacts than the previously used site homogeneous models (Lartillot and Philippe, 2004). All these results argue in favor of a deep branching of group A at the base of Haloarchaea.

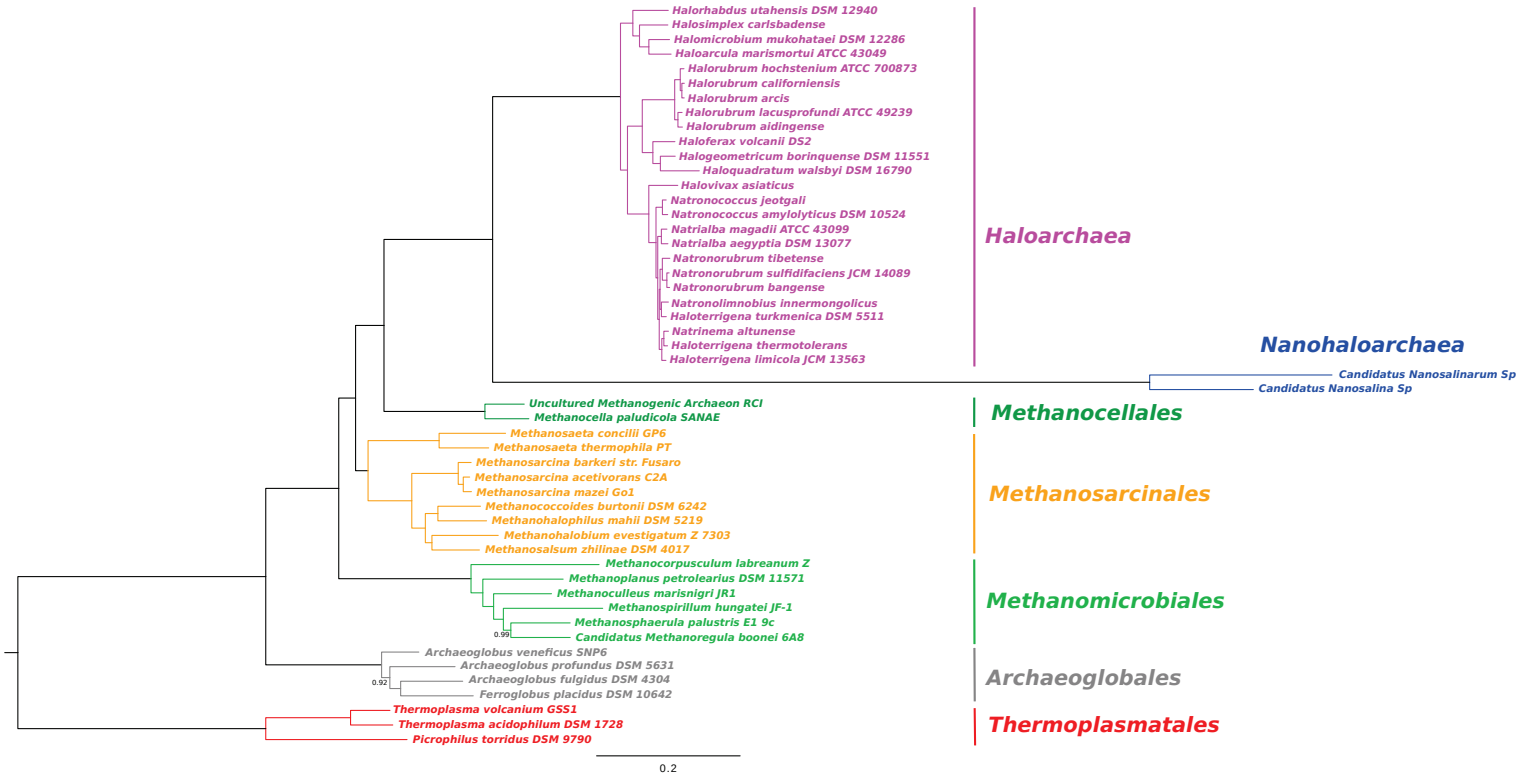



Figure 3: **Species tree reconstruction with the CAT model**. This consensus posterior tree was reconstructed with the CAT model in Phylobayes. All branches with no support information have a PP of 1. The topology within Haloarchaea is identical to the ML topology (Figure 1).

## MalDH joint gene tree

The MalDH sequence-only tree inferred with the CAT-GTR model is globally weakly supported in most regions of the Haloarchaea tree (Figure 4).

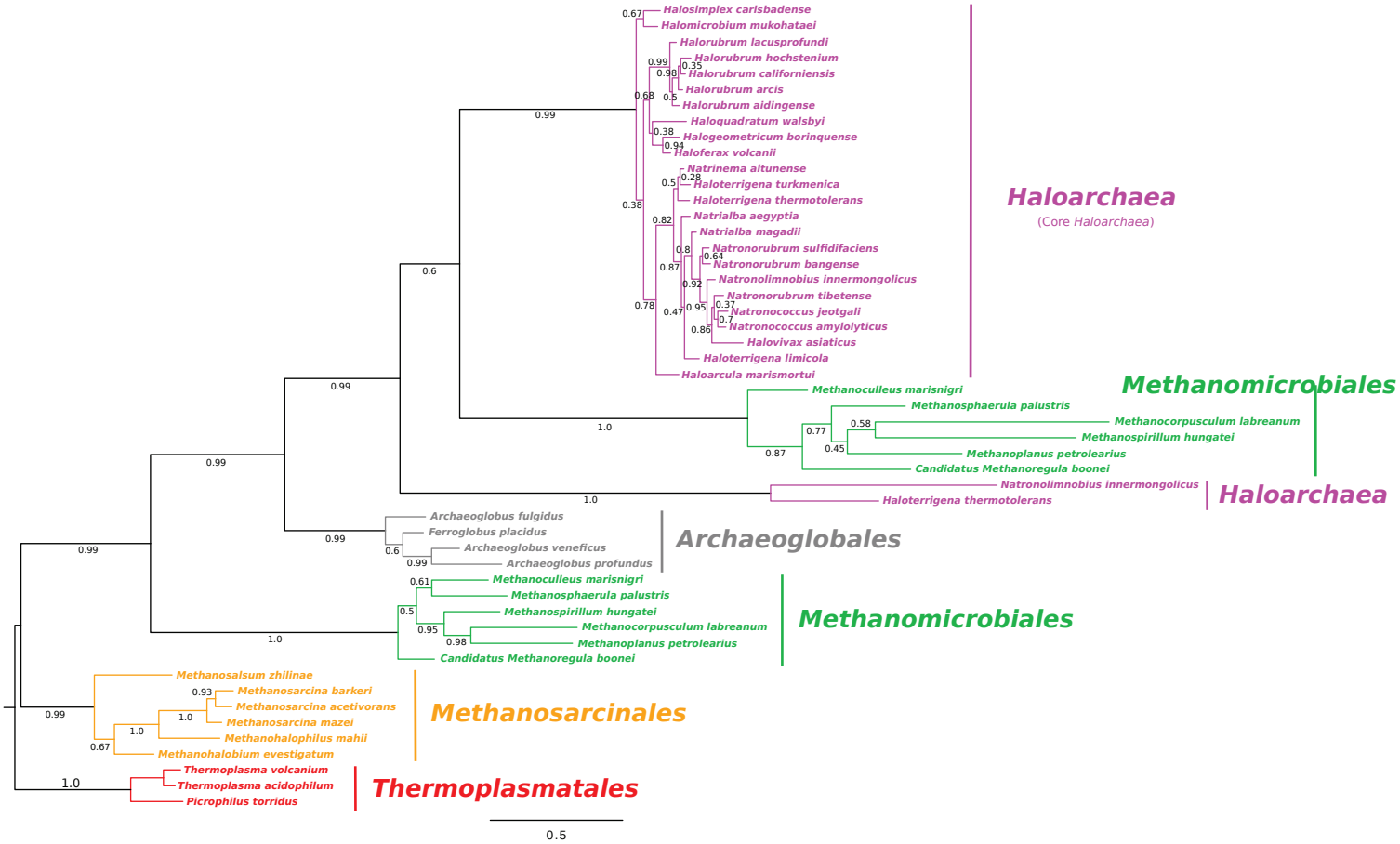



Figure 4: **Sequence-only tree of MalDH sequences**. The tree was inferred with Phylobayes and the CAT-GTR model. Supports on branches are posterior probabilities. Phyla are color-coded in the same way as in Figure 1.

Two haloarchaeal species, *Haloterrigena thermotolerans* and *Natronolimnobius innermongolicus* contain two copies of MalDH in their genome. These two second copies are very divergent from the other haloarchaeal sequences and branch deeply in the tree, at the base of one of the two copies possessed by Methanomicrobiales, making Haloarchaea paraphyletic. Nonetheless, the support for this position is weak (PP=0.6), probably due to the long branch leading to these two sequences. Regarding the core part of the Haloarchaea tree, the Robinson-Foulds distance between the topology of the sequence-only tree and that of the species tree equals 28, which means that, on average, 2/3 of the bipartitions present in one of the two trees are not present in the other. The high topological distance and the presence of weakly-supported nodes highlights the need to perform a probabilistic reconciliation to infer the most likely scenario of transfers, duplications and losses while taking into account the uncertainty in the sequence-only tree resolution. We used ALE (Szöllősi et al., 2013a), which implements the exODT model (Szöllősi et al., 2013b) (see Material and Methods) to compute the ML joint tree, represented in Figure 5.

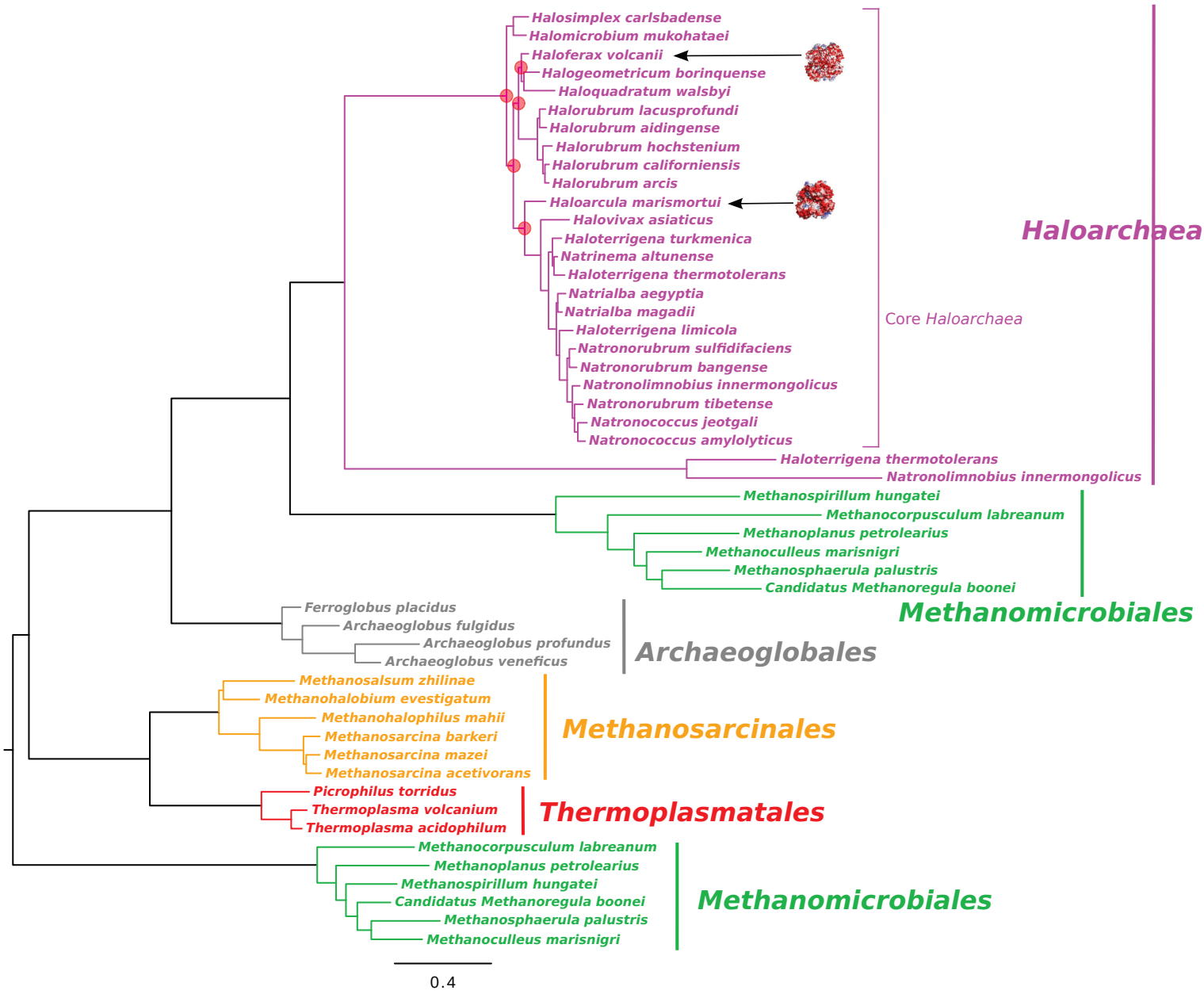



Figure 5: **Joint tree of MalDH**. The phylogenetic tree maximizing the joint likelihood of sequence and DTL events information was calculated with ALE. This tree was used to reconstruct ancestral sequences of extant MalDHs. Protein models indicate ancestors for which a resurrection was performed. Cristal structures show species for which a 3D structure has been determined (*H. marismortui* and *H. volcanii*). Core Haloarchaea are all haloarchaeal sequences excepted the two MalDH duplicates present in *Haloterrigena thermotolerans* and *Natronolimnobius innermongolicus*. Phyla are color-coded as in Figure 1.

It appears that the two extreme copies of MalDH sequences in *Haloterrigena thermotolerans* and *Natronolimnobius innermongolicus* are now placed at the base of Haloarchaea. However, the sequence information strongly rejects a placement of these sequences within core Haloarchaea. The MalDH joint tree of core Haloarchaea presented in Figure 5 is now much closer to the species tree, as attested by the lower Robinson-Foulds distance, now equals to 16. It shows that ALE detected that a large part of the topological inconsistency was due to phylogenetic uncertainty present in the MalDH sequences.

Figure 6 represents the ML reconciliation scenario in terms of duplication, transfer and loss of gene events in Haloarchaea inferred by ALE. The deep branching of the two extreme duplicate copies of MalDH is associated to a deep transfer event outside the species tree, that subsequently came back to an ancestral lineage of *Haloterrigena thermotolerans* and *Natronolimnobius innermongolicus*. In core Haloarchaea, 1 duplication, 3 transfers and 14 losses are inferred.

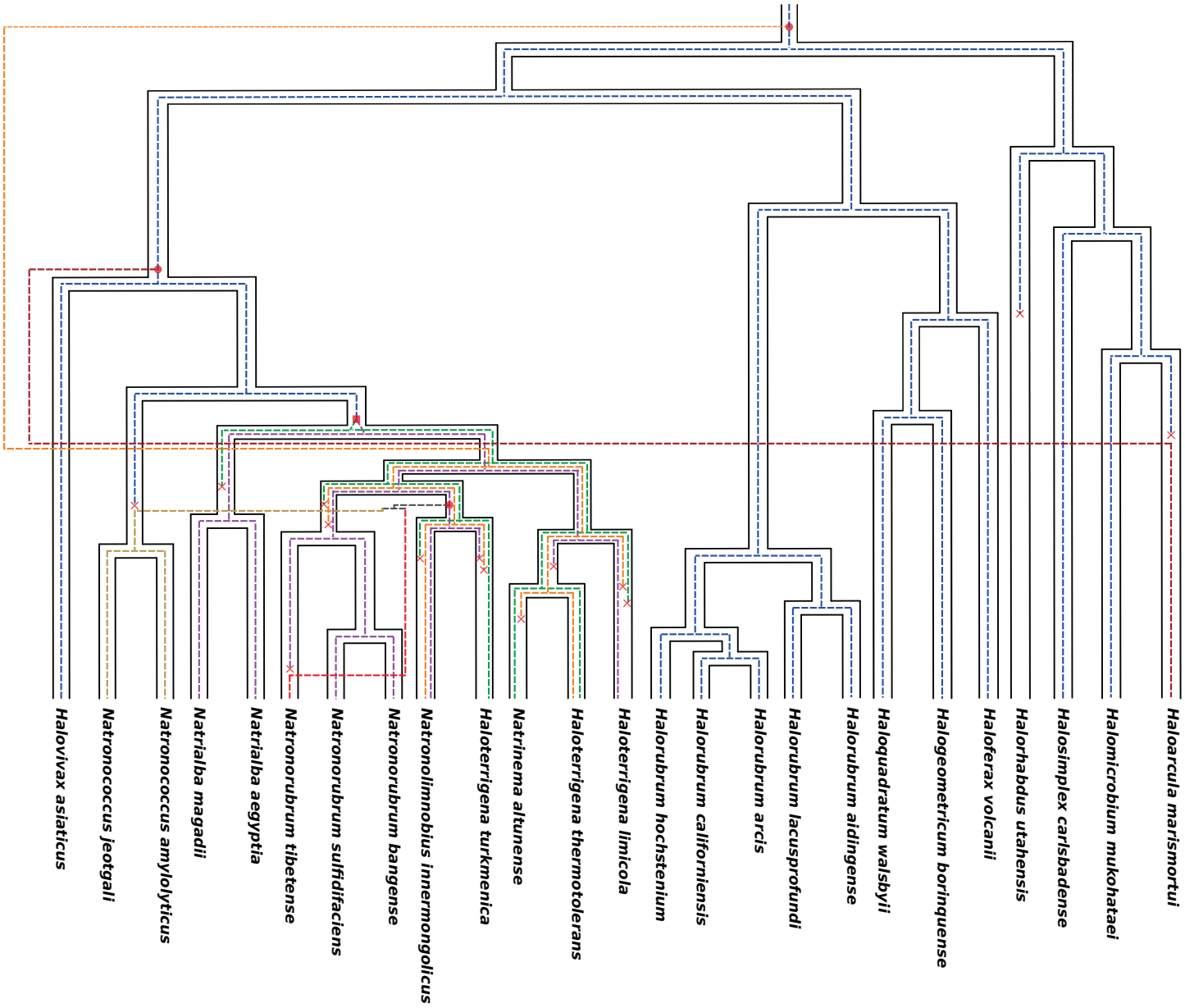



Figure 6: **MalDH gene family history**. This figure represents the evolution of MalDH sequences in terms of duplications, transfers and losses, embedded in the species tree of Haloarchaea. The red square represents a duplication event. Red circles represent transfer events. Red crosses represent loss events. Vertical evolution of MalDH sequences outside the species tree represents the evolution of the sequence in unsampled or extinct species.

Three of these losses are associated to a transfer event, so that the gene loss may represent a non-homologous recombination between the ancient and the transferred MalDH copy. Figure 6 shows that

| Process | Model | ln(Likelihood) | Number of Parameters | AIC | BIC |
|---|---|---|---|---|---|
| Site- and Time-homogeneous | LG | -15129.2 | 1 | 30260.4 | 30270.3 |
| | LG+$F_{opt}$ | -15075.4 | 20 | 30190.8 | 30269.81 |
| Site-heterogeneous | UL2 | -15011 | 2 | 30026 | 30033.9 |
| | **UL3** | **-14963.6** | **3** | **29933.2** | **29945.05** |
| | EX2 | -14999.3 | 2 | 30002.6 | 30010.5 |
| | EX3 | -14970.1 | 3 | 29946.2 | 29958.05 |
| | EHO | -15022.6 | 3 | 30051.2 | 30063.05 |
| | C10 | -15433.6 | 10 | 30887.2 | 30926.71 |
| | C20 | -15357.6 | 20 | 30755.2 | 30834.21 |
| | C30 | -15307.1 | 30 | 30674.2 | 30792.72 |
| | C40 | -15243.6 | 40 | 30567.2 | 30725.23 |
| | C50 | -15252.4 | 50 | 30604.8 | 30802.33 |
| | C60 | -15229.3 | 60 | 30578.6 | 30815.64 |
| Time-heterogeneous | LG+COaLA[1] | -15009.2 | 103 | 30224.4 | 30631.32 |
| | LG+COaLA[2] | -14928.9 | 204 | 30265.8 | 31071.73 |

Table 3: **Selection of the best-fitting model to reconstruct ML ancestral sequences of MalDH**. LG+COaLA[$k$] indicates that the LG exchangeability matrix was considered when using the COaLA model, with $k$ branch-specific parameters.

the MalDH gene family history is complex and that the reconciliation approach is required to have an accurate representation of the gene diversification pattern, a major prerequisite to subsequently perform ancestral sequence reconstructions (Groussin et al., 2013c). However, this ML reconciliation scenario is much simpler than a reconciliation scenario computed with the MalDH sequence-only tree that would ignore the phylogenetic uncertainty of the MalDH sequence-only tree. We used Mowgli (Nguyen et al., 2013) to compute this maximum parsimony scenario for core Haloarchaea. Mowgli inferred 0 duplication, 8 losses and up to 10 transfer events, underlining the need to take sequence information into account to refine trees and to propose more reasonable gene family histories.

## Inference of ancestral MalDH sequences

Groussin et al. (2013c) emphazised the need to use complex evolutionary models such as site- or time-heterogeneous models to infer accurate ancestral sequences. Model fit was shown to strongly correlate with ASR accuracy, such that the model having the best fit measured with model selection criteria should be used to perform ASR (Groussin et al., 2013c). Model fit is straightforward to measure in the Maximum Likelihood framework, by using AIC or BIC criteria that attribute a penalty to the final likelihood depending on the number of parameters estimated by the model. Table 3 shows that, among all models tested, the site-heterogeneous UL3 model (Le et al., 2008b) is the best at fitting the data according to both AIC and BIC criteria. This model was chosen to perform ASR of the MalDH sequences in ML. Within the Bayesian framework, we used the CAT-GTR model, which outperforms its counterparts CAT, and GTR (not shown).

Globally, ancestral sequences inferred in core Haloarchaea have strong support according to the

posterior probabilities of the ancestral ML states. Thus, 95%, 95%, 94%, 96% and 96% of ancestral residues have PP higher than 0.9 for ANC1, ANC2, ANC3, ANC4 and ANC5 respectively (See Figure 7 for the labelling of ancestral sequences). Ancestral sequences obtained with CAT-GTR are slightly less supported, as 94%, 94%, 93%, 92% and 92% of residues have PP $>$ 0.9, respectively. However, further experiments are needed to conclude if this weaker support is due to the difference between ML vs. Bayesian reconstruction or to the difference between evolutionary models.

For the five targeted ancestors, we compared ancestral sequences inferred with UL3 and CAT-GTR. It appears that 5, 3, 4, 3 and 6 amino acids differ between the two models for ANC1, ANC2, ANC3, ANC4 and ANC5 respectively. It only represents 1 to 2% of total sites, showing that for this particular dataset, UL3 and CAT-GTR are able to produce very similar sequences. Some of these differences concern amino acids with very similar biochemical properties (L$\leftrightarrow$I or A$\leftrightarrow$V for instance). However, 7 of these changes imply amino acids that are biochemically different, such as V$\leftrightarrow$T, P$\leftrightarrow$Q or S$\leftrightarrow$R. Since these latter differences are more likely to have an impact on protein catalysis and structure, further control experiments are needed to clarify if these residues may have an impact on catalysis, intra- or inter-monomer interactions and solvent interactions. It is worth noting that even the replacement of a given amino acid by another one sharing similar biochemical properties may have a short- or long-distance impact on the conformational landscape by modifying its steric and/or electronic environment, which is a key aspect in the adaptation of halophilic proteins to high salt concentrations. Over the 21 differences between UL3 and CAT-GTR for the 5 ancestors, 9 concern sites that are poorly- (PP $<$ 0.7) to weakly-supported (0.7 $<$ PP $<$ 0.8) according to PP computed with UL3. This shows that a substantial part of these differences lies in a genuine difficulty to accurately decipher the nature of the ancestral state whatever the substitution model employed. However, 10 of these differences concern sites that are mildly-supported (0.8 $<$ PP $<$ 0.9, some of them have PP equal to 0.87 or 0.89). It confirms the influence of substitution models on ASR inference such that strongly supported ancestral residues obtained with a given model may become poorly supported with another one (Groussin et al., 2013c). In line with this, among the 7 diffences between UL3 and CAT-GTR implying amino acids that are biochemically different, 4 of them are mildly- to strongly-supported (PP $>$ 0.9) by UL3.

Results presented in Figure 7 show the substitutional pathway of MalDH from the last common ancestor of core Haloarchaea to *H. marismortui* and *H. volcanii*. Given the crystallographic structure of Hma MalDH, structure predictions were performed for ANC1 to ANC5 (data not shown). It allowed us to observe that several of these substitutions concern residues that are involved in inter-monomer interactions and so, may have strong influence on regulating protein stability.

It appeared that the ancestral MalDH sequences of *H. marismortui* and *H. volcanii* reconstructed with the sequence-only tree and the joint tree differ by only four amino acids. Consequently, the impact of the phylogenetic tree is in the same range than the impact of the substitution model (see above). It

confirms that ancestors of MalDHs in Haloarchaea can be inferred with reasonable confidence.

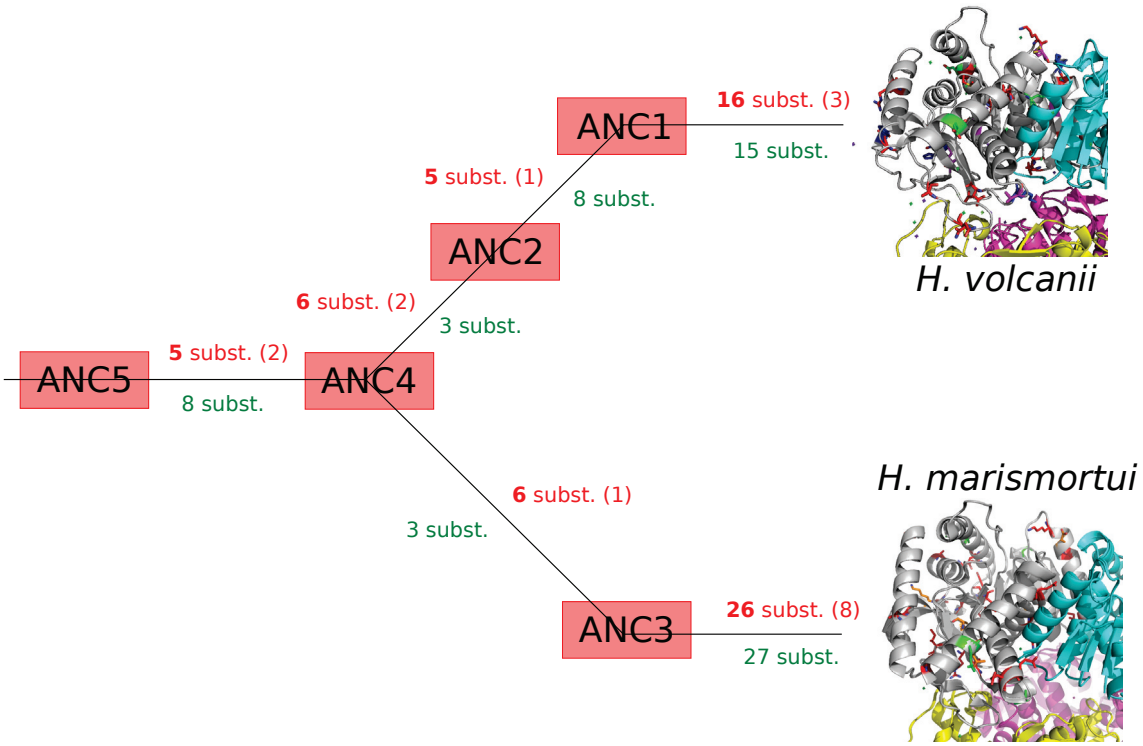



Figure 7: **Substitution history from the last common ancestor of core Haloarchaea to *H. marismortui* and *H. volcanii*.** ANC1 to ANC5 represent the 5 target ancestors that were experimentally resurrected. The number of substitutions between sequences inferred with CAT-GTR is indicated in red. Numbers between brackets refer to substitutions involving residues at the interface between monomers. The number of substitutions inferred with UL3 is indicated in green.

## Resurrection of ancestral MalDH

We first compared the stability properties of contemporary MalDH enzymes of *H. marismortui* and *H. volcanii*. Figure 8 shows that Hma MalDH requires high salt concentrations to be active as it is in its native state for a KCl concentration ranging from 1.8M to 4M. Though displaying high sequence identity (80%) with Hma MalDH, Hvo MalDH has a higher conformational stability, as it is very stable even at low KCl concentrations (0.5M). This difference directly reflects differential ecological adaptations, as *H. marismortui* lives in high salt environments. However, at very low ion concentration these two enzymes are unfolded, showing that they are clearly halophilic enzymes which require a sufficient ion concentration to be stabilized and active. It highlights the role of cations and anions in the tertiary and quaternary stability of MalDH.

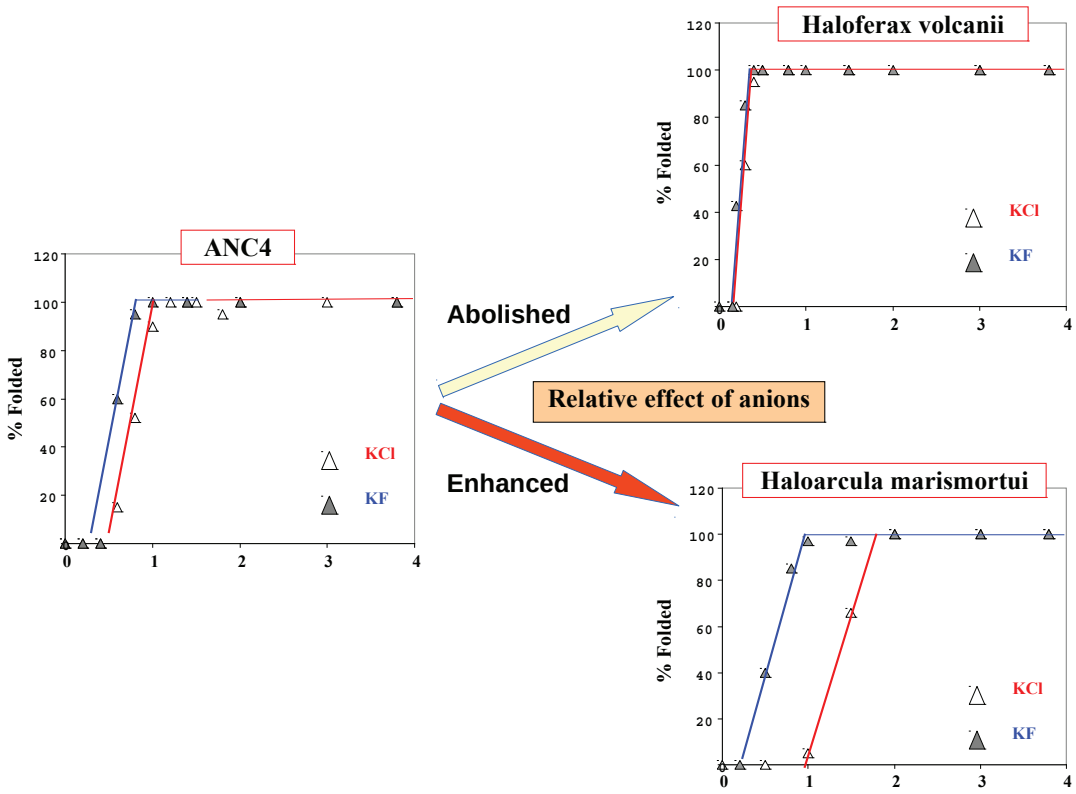



Figure 8: **Salt-dependent stability of extant and ancestral MalDH**. Only results for ANC4 are provided. Stability was measured with .... Effect of anions on stability was measured with a constant cation ($K^+$) and two different anions ($Cl^-$, white triangles and $F^-$, grey triangles).

The same experiment was realised by replacing the chloride anion with the fluoride ($F^-$) anion, which has a higher electronegativity. $F^-$ is not naturally present in the cytoplasm of halophilic organisms. However, it allows to decipher the influence of ions (here of anions) in the stability of the MalDH structure. In the presence of KF, Hma MalDH is stable from lower ion concentrations (about 1M) than with KCl, showing that it has a much higher conformational stability with fluoride. This demonstrates that interactions involving MalDH residues and ions of the solvent are stabilised when ions of high electronegativity are used. This also highlights the crucial role of the MalDH residues involved in these interactions that strongly determine the stability of the protein. Concerning *H. volcanii*, the stabilizing effect of the $F^-$ anion is abolished (Figure 8). This underlines that intra or inter-subunit interactions play a more important role in the stabilisation of the quaternary structure of *H. volcanii* than of *H. marismortui*.

Stability curves for ANC4 are shown in Figure 8. Despite the high global similarity between ANC4 and Hma and Hvo MalDHs at the sequence level (Figure 7), ANC4 displays different stability properties. In the presence of KCl, ANC4 is more stable than Hma MalDH. This may suggest that this ancestor required lower ion concentration in the cytoplasm than did *H. marismortui*. However, such conclusions should be drawn with caution (see Discussion). What is interesting is that fluoride has an intermediate

influence on ANC4 MalDH in terms of stabilizing effects due to interactions with anions. These effects were enhanced in *H. marismortui* and abolished in *H. volcanii* (Figure 8). It is probably linked to a variation of affinity for anions with respect to residues involved in solvent interactions.

## Discussion

With the present study, we aim at using ASR and resurrection of ancient halophilic MalDH proteins to provide the first detailed description of the influence of the successive amino-acid substitutions in the shaping of the conformational landscape of MalDH during adaptation to extreme salt concentrations. ASR is performed along a phylogenetic tree whose the reconstruction needs careful attention, as it was shown that topology incongruences owing to DTL and sequence uncertainty may strongly influence ASR accuracy (Groussin et al., 2013c).

The ALE model, by considering both sequence and species information in the calculation of the ML joint reconciled tree, proposed a reconciliation scenario that is much more parsimonious than the one inferred without accounting for sequence uncertainty. This shows the importance of jointly considering the different sources of topological incongruences in the computation of reconciled trees. In the ML joint tree, the two halophilic paralogs branch at the base of core Haloarchaea and not within. This branching is associated to an ancient horizontal gene transfer to unsampled or extinct lineages at the base of the core Haloarchaea clade that later came back to ancestral lineages of extant core Haloarchaea. On a general basis, two reasons may explain why the two paralogs were not placed within core Haloarchaea. The first reason is that no trees having these two copies branching within core haloarchaeal sequences were present in the sample of posterior sequence-only trees given to ALE, so that ALE could not test a reconciliation with the species tree. The second less likely reason is that sequence information penalizes too much the joint likelihood computed by ALE. Indeed, a clade containing the two paralogs branching within core Haloarchaea species may have been present in the posterior sample of sequence-only trees, such that ALE was able to compute a joint likelihood with a perhaps more likely reconciliation scenario. But the gain in likelihood due to this more likely scenario was not enough to compensate the stronger likelihood penalty brought by sequence information owing to the branching of the two paralogs. Here, about 0.002 trees (210 over 95242 trees present in the posterior sample) contain the two paralogs within core Haloarchaea. It shows that sequence information reject too strongly a close relationship between the two paralogs and other Haloarchaea sequences to be placed within the core species in the joint tree.

Transfer events are inferred conditional on the divergence times of the species tree. These divergence times were calculated with a relaxed-clock model and only the phylogenetic signal present in the concatenate of genes, as no fossil calibration in Archaea is available. Consequently, these dates necessarily contain uncertainty that should impact reconciliation scenarios (but not the detection of transfers). In

conclusion, the ML scenario proposed by ALE should be interpreted with great care.

In addition to the first biochemical characterization performed on ANC4 in terms of stability, further experiments are needed to fully characterize the discriminative effects of the various amino acids substitutions on MalDH properties. For instance, similar investigations on the role of anions and cations in protein-solvent interactions are needed for the other ANC1, ANC2, ANC3 and ANC5 ancestors. They should provide information on the range of salt concentrations to which ancient MalDH were adapted. These results are necessary to predict ancient environmental conditions in which ancestors of extant halophiles lived. Furthermore, salt-dependent stability measurements will allow to measure the relative role of ion-binding sites on MalDH stability. In the present study, only the role of anions was investigated, by replacing $Cl^-$ by $F^-$ while conserving $K^+$. However, it is also knwon that the magnesium cation ($Mg^{2+}$) is able to regulate the stability properties of MalDH (Madern and Zaccaï, 1997; Ebel et al., 1999). The determination of X-ray crystallographic structures for each of the ANC proteins will give clues to comprehend the unpredictable long-range effects of substitutions on the stability properties of proteins, especially on electrostatic environment of ion-binding sites. For the moment, only *in-silico* 3D structure predictions were realized using the Hma MalDH structure and the SWISS-MODEL program (Arnold et al., 2006; Kiefer et al., 2009). These preliminary results are promising because they allowed us to identify what particular substitutions should be involved in intra- or inter-monomer or protein-solvent interactions and could potentially influence protein stability. However, these 3D models are uninformative with respect to the positioning of anions and cations that are in interactions with ion-binding sites. Only X-ray crystallographic structures will provide this information.

Prediction of ancient environments from the activity and stability data obtained *in-vitro* on resurrected single proteins should be attempted with great care. It especially concerns our case, as *in-vitro* media does not precisely represent the composition of ions in the cytoplasm. In halophiles, different types of anions and cations regulate protein stability. The stability experiment performed in this study is unidimensional as it only tests the influence of one ion composition at a time and so, may not provide an accurate protein phenotype in terms of stability. One way to overcome this potential bias may be to use the relation between adaptation to salt concentrations and isoelectric point of proteomes. The isoelectric point of halophilic organisms has been shown to be particularly acidic, due to global enrichment of acidic residues in halophilic proteomes (Paul et al., 2008). Kiraga et al. (2007) have shown that the isoelectric point computed at the scale of proteomes was able to discriminate between extreme, moderate and non halophiles. If a significant correlation between salinity conditions of life of each haloarchaea present in this study and their global isoelectric point is found, a linear regression model could be used to predict ancient salinities of life over the tree of Haloarchaea from ancestral proteome compositions reconstructed with non-homogeneous substitution models (Boussau et al., 2008; Groussin and Gouy, 2011). Further investigations are needed to test the feasability of such an approach and

to compare these inferences with predictions made from *in-vitro* stability measurements performed on ancestral MalDH.

# References


Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* pages 267–281. Petrov BN, Csaki F, editors Budapest (Hungary).

Åkerborg O, Sennblad B, Arvestad L, and Lagergren J. 2009. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc Natl Acad Sci U S A* 106:5714–5719.

Arnold K, Bordoli L, Kopp J, and Schwede T. 2006. The SWISS-MODEL Workspace: A web-based environment for protein structure homology modelling. *Bioinformatics* 22:195–201.

Baliga NS, Bonneau R, Facciotti MT, et al.. 2004. Genome sequence of Haloarcula marismortui: A halophilic archaeon from the Dead Sea. *Genome Res* 14:2221–2234.

Birktoft JJ, Feruley RT, Bradshaw RA, Banasazk LJ, et al.. 1982. Amino acid sequence homology among the 2-hydroxy acid dehydrogenases: mitochondrial and cytoplasmic malate dehydrogenases from a mologous system with lactate dehydrogenase. *Proc Natl Acad Sci U S A* 79:6166–6170.

Blanquart S and Lartillot N. 2006. A Bayesian Compound Stochastic Process for Modeling Nonstationary and Nonhomogeneous Sequence Evolution. *Mol Biol Evol* 23:2058–2071.

Boussau B, Blanquart S, Necsulea A, Lartillot N, and Gouy M. 2008. Parallel Adaptation to High Temperature in the Archaean Eon. *Nature* 456:942–945.

Boussau B, Szöllősi GJ, Duret L, Gouy M, Tannier E, and Daubin V. 2013. Genome-scale coestimation of species and gene trees. *Genome Res* 23:323–330.

Brinkmann H, van der Giezen M, Zhou Y, de Raucourt GP, and Philippe H. 2005. An empirical Assessment of Long Branch Attraction Artefacts in Deep Eukaryotic Phylogenomics.pdf. *Syst Biol* 54:743–757.

Brochier-Armanet C, Boussau B, Gribaldo S, and Forterre P. 2008. Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nat Rev Microb* 6:245–252.

Brochier-Armanet C, Forterre P, and Gribaldo S. 2011. Phylogeny and evolution of the Archaea: one hundred genomes later. *Curr Op Microb* 14:274–281.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17:540–552.

Chang B, Jönsson K, Kazmi M, Donoghue MJ, and Sakmar TP. 2002. Recreating a Functional Ancestral Archosaur Visual Pigment. *Mol Biol Evol* 19(9):1483–1489.

Dill KA and Chan HS. 1997. From Levinthal to pathways to funnels. *Nat Struct Biol* 4:10–19.

Dutheil J and Boussau B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol Biol* 8:255.

Dym O, Mevarech M, Sussman JL, et al.. 1995. Structural Features That Stabilize Halophilic Malate Dehydrogenase from an Archaebacterium. *Science* 267:1344–1346.

Ebel C, Faou P, Kernel B, and Zaccaï G. 1999. Relative role of anions and cations in the stabilization of halophilic malate dehydrogenase. *Biochemistry* 38:9039–9047.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.

Finnigan GC, Hanson-Smith V, Stevens TH, and Thornton JW. 2012. Evolution of increased complexity in a molecular machine. *Nature* 481:360–364.

Foster PG. 2004. Modeling compositional heterogeneity. *Syst Biol* 53:485–495.

Frauenfelder H, Chen G, Berendzen J, Fenimore PW, Jansson H, McMahon BH, Stroe IR, Swenson J, and Young RD. 2009. A unified model of protein dynamics. *Proc Natl Acad Sci U S A* 106:5129–5134.

Gaucher EA, Thomson JM, Burgan MF, and Benner SA. 2003. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature* 425:285–288.

Groussin M, Boussau B, and Gouy M. 2013a. A Branch-Heterogeneous Model of Protein Evolution for Efficient Inference of Ancestral Sequences. *Syst Biol* 62:523–538.

Groussin M and Gouy M. 2011. Adaptation to Environmental Temperature Is a Major Determinant of Molecular Evolutionary Rates in Archaea. *Mol Biol Evol* 28:2661–2674.

Groussin M, Guéguen L, Boussau B, Gouy M, and Lartillot N. 2013b. Efficient modeling of protein site-heterogeneities with empirical mixtures of profiles. *Chapitre 2, section 2 de ce manuscrit de thèse.*

Groussin M, Hobbs JK, Szöllősi GJ, Gribaldo S, Arcus VL, and Gouy M. 2013c. Biologically motivated models strongly improve the functionality of resurrected proteins. *Chapitre 4, section 1 de ce manuscrit de thèse.*

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, and Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59:307–321.

Guindon S and Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704.

Harms MJ and Thornton JW. 2010. Analyzing protein structure and function using ancestral gene reconstruction. *Curr Opin Struct Biol* 20:360–366.

Hobbs JK, Shepherd C, Saul DJ, Demetras NJ, Haaning S, Monk CR, Daniel RM, and Arcus VL. 2012. On the Origin and Evolution of Thermophily: Reconstruction of Functional Precambrian Enzymes from Ancestors of Bacillus. *Mol Biol Evol* 29:825–835.

Höhna S and Drummond AJ. 2012. Guided tree topology proposals for bayesian phylogenetic inference. *Syst Biol* 61:1–11.

Hrdy I, Hirt R, Dolezal P, Bardanova L, Foster P, Tachezy J, and Embley T. 2004. Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. *Nature* 432:618–622.

Irimia A, Ebel C, Madern D, Richard SB, Cosenza LW, Zaccaï G, and Vellieux FMD. 2003. The oligomeric states of Haloarcula marismortui malate dehydrogenase are modulated by solvent components as shown by crystallographic and biochemical studies. *J Mol Biol* 326:859–873.

Jaenicke R. 2000. Stability and stabilization of globular proteins in solution. *J Biotechnol* 79:193–203.

Kiefer F, Arnold K, Künzli M, Bordoli L, and Schwede T. 2009. The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res* 13:658–660.

Kiraga J, Mackiewicz P, Mackiewicz D, Kowalczuk M, Biecek P, Polak N, Smolarczyk K, Dudek MR, and Cebrat S. 2007. The relationships between the isoelectric point and: length of proteins, taxonomy and ecology of organisms. *BMC Genomics* 8:163.

Kodra JT, Skovgaard M, Madsen D, and Liberles DA. 2007. Linking sequence to function in drug design with ancestral sequence reconstruction. In *Ancestral Sequence Reconstruction* pages 34–39. Oxford University Press.

Lartillot N, Lepage T, and Blanquart S. 2009. PhyloBayes 3. A Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25:2286–2288.

Lartillot N and Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol* 21:1095–2004.

Le SQ and Gascuel O. 2008. An Improved General Amino Acid Replacement Matrix. *Mol Biol Evol* 25:1307–1320.

Le SQ and Gascuel O. 2010. Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. *Syst Biol* 59:277–287.

Le SQ, Gascuel O, and Lartillot N. 2008a. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24:2317–2323.

Le SQ, Lartillot N, and Gascuel O. 2008b. Phylogenetic mixture models for proteins. *Phil. Trans. R. Soc. Lond. B* 363:3965–3976.

Lee D, Redfern O, and Orengo C. 2007. Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol* 8:995–1005.

Liu Y and Bahar I. 2012. Sequence Evolution Correlates with Structural Dynamics. *Mol Biol Evol*.

López-García P, Moreira D, López-López A, and Rodríguez-Valera F. 2001. A novel haloarchaeal-related lineage is widely distributed in deep oceanic regions. *Environ Microbiol* 3:72–78.

Löytynoja A and Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci U S A* 102:10557–10562.

Löytynoja A and Goldman N. 2008. Phylogeny-Aware Gap Placement Prevents Errors in Sequence Alignment and Evolutionary Analysis. *Science* 320:1632–1635.

Madern D, Ebel C, Dale HA, Lien T, Steen IH, Birkeland NK, Zaccai G, et al.. 2001. Differences in the oligomeric states of the [LDH-like] L-MalDH from the hyperthermophilic archaea Methanococcus jannaschii and Archaeoglobus fulgidus. *Biochemistry* 40:10310–10316.

Madern D, Ebel C, et al.. 2007. Influence of an anion-binding site in the stabilization of halophilic malate dehydrogenase from Haloarcula marismortui. *Biochimie* 89:981–987.

Madern D, Ebel C, and Zaccai G. 2000. Halophilic adaptation of enzymes. *Extremophiles* 4:91–98.

Madern D and Zaccaï G. 1997. Stabilisation of halophilic malate dehydrogenase from haloarcula marismortui by divalent cations. *Eur J Biochem* 249:607–611.

Malcolm B, Wilson K, Matthews B, Kirsch J, and Wilson A. 1990. Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. *Nature* 345:86–89.

Miele V, Penel S, and Duret L. 2011. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics* 12:116.

Mirceta S, Signore A, Burns J, Cossins A, Campbell K, and Berenbrink M. 2013. Evolution of Mammalian Diving Capacity Traced by Myoglobin Net Surface Charge. *Science* 340(6138).

Narasingarao P, Podell S, Ugalde JA, Brochier-Armanet C, Emerson JB, Brocks JJ, Heidelberg KB, Banfield JF, and Allen EA. 2012. De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. *ISME J* 6:81–93.

Nguyen TH, Ranwez V, Pointet S, Chifolleau AMA, Doyon JP, and Berry V. 2013. Reconciliation and local gene tree rearrangement can be of mutual profit. *Algorithms for Molecular Biology* 8:12.

Paul S, Bag SK, Das S, Harvill ET, and Dutta C. 2008. Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. *Genome Biol* 9:R70.

Penn O, Privman E, Landan G, Graur D, and Pupko T. 2010. An alignment confidence score capturing robustness to guide-tree uncertainty. *Mol Biol Evol* 27:1759–1767.

Philippe H, Delsuc F, Brinkmann H, and Lartillot N. 2005. Phylogenomics. *Annu Rev Ecol Evol Syst* 36:541–562.

Rasmussen MD and Kellis M. 2012. Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Res* 22:755–765.

Richard SB, Madern D, Garcin E, and Zaccai G. 2000. Halophilic Adaptation: Novel Solvent Protein Interactions Observed in the 2.9 and 2.6 Å Resolution Structures of the Wild Type and a Mutant of Malate Dehydrogenase from Haloarcula marismortui. *Biochemistry* 39:992–1000.

Rodriguez-Ezpeleta N, Brinkmann H, Roure B, and Lartillot N. 2007. Detecting and Overcoming Systematic Errors in Genome-Scale Phylogenies. *Syst Biol* 56:389–399.

Schwarz G. 1978. Estimating the Dimension of a Model. *Ann Statist* 6:461–464.

Shimodaira H and Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17:1246–1247.

Stackhouse J, Presnell S, McGeehan G, Nambiar K, and Benner S. 1990. The ribonuclease from an extinct bovid ruminant. *FEBS Lett* 262:104–106.

Sundaram TK, Wright IP, Wilkinson AE, et al.. 1980. Malate dehydrogenase from thermophilic and mesophilic bacteria. Molecular size, subunit structure, amino acid composition, immunochemical homology and catalytic activity. *Biochemistry* 19:2017–2022.

Szöllősi GJ, Rosikiewicz W, Boussau B, Tannier E, and Daubin V. 2013a. Efficient Exploration of the Space of Reconciled Gene Trees. *Syst Biol*.

Szöllősi GJ, Tannier E, Lartillot N, and Daubin V. 2013b. Lateral Gene Transfer from the Dead. *Syst Biol* 62:386–397.

Tokuriki N and Tawfik DS. 2009. Protein dynamism and evolvability. *Science* 324:203–207.

Voordeckers K, Brown CA, Vanneste K, van der Zande E, Voet A, Maere S, and Verstrepen KJ. 2012. Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. *PLoS Biol.* 10(12):e1001446.

Williams TA, Foster PG, Nye TMW, Cox CJ, and Embley TM. 2012. A congruent phylogenetic signal places eukaryotes within the Archaea. *Proc Biol Sci* 279:4870–4879.

Wu YC, Rasmussen MD, Bansal MS, and Kellis M. 2013. TreeFix: Statistically Informed Gene Tree Error Correction Using Species Trees. *Syst Biol* 62:110–120.

Yang Z, Kumar S, and Nei M. 1995. A New Method of Inference of Ancestral Nucleotide and Amino Acid Sequences. *Genetics* 141:1641–1650.

Zaccai G. 2004. The effect of water on protein dynamics. *Phil. Trans. R. Soc. Lond. B* 359:1269–1275.

# 5 Perspectives

J'envisage de nombreuses perspectives à ce travail de thèse. J'en propose quelques unes ci-dessous en espérant avoir la possibilité de les aborder dans les années futures.

Tout d'abord, l'amélioration de la compréhension du monde vivant et des processus évolutifs agissant au niveau des génomes passera par l'amélioration des modèles de substitution et de reconstruction des arbres phylogénétiques. La réalité de l'évolution des séquences est beaucoup plus complexe que ce que les modèles actuels sont capable de prendre en compte, y compris les plus complexes. Malgré cela, cette complexité ne doit pas être un frein aux développements méthodologiques et ne doit pas être une raison pour abandonner d'essayer de comprendre l'évolution du vivant par des méthodes phylogénétiques. Suite aux travaux présentés dans cette thèse, j'envisage de continuer à améliorer les modèles permettant de mieux prendre en compte les mécanismes évolutifs, tels que les variations de processus et/ou taux entre lignées et entre sites. Ainsi, le modèle hétérogène à la fois en temps et en sites présenté dans la section 2.3 pourrait permettre de modéliser efficacement en ML ces hétérogénéités, ouvrant la voie à la résolution de questions phylogénétiques difficiles.

Contrairement aux séquences nucléiques, il semble, d'après des tests effectués durant cette thèse et des discussions avec Nicolas Lartillot, que les biais de compositions globales n'affectent

pas fortement les reconstructions phylogénétiques en acides aminés. Le modèle COaLA présenté en section 2.1 permet de prendre en compte ces variations globales. Seulement, il n'est pour le moment pas possible de réaliser une exploration des topologies avec COaLA, qui ne fonctionne qu'avec une topologie fixe. Il serait intéressant d'intégrer ce modèle au programme PhyML (Guindon et al., 2010), qui permet de réaliser des explorations de l'espace des topologies très efficaces en termes de temps de calcul. Une collaboration a été entamée avec Bastien Boussau et Sebastien Höhna pour l'utilisation de COaLA dans RevBayes, qui permet de réaliser des analyses phylogénétiques en Bayesien. Dernièrement, le modèle a été implémenté dans les codes de RevBayes. Il va être très intéressant de tester si COaLA permet d'améliorer les reconstructions phylogénétiques lorsque les compositions moléculaires globales varient d'une espèce à l'autre, sur données simulées et données réelles.

Il serait également intéressant de savoir quelles sont les capacités réelles des modèles de substitution hétérogènes en temps à correctement inférer la position de la racine d'un arbre phylogénétique. L'intérêt majeur est d'éviter d'avoir recours à un groupe externe le plus souvent distant des espèces du groupe interne, créant de gros problèmes de reconstruction la plupart du temps entrainés par le biais d'attraction des longues branches. Il n'est pas évident de savoir *a priori* si le signal phylogénétique lié à la variation des compositions dans le temps soit suffisant pour discriminer une position de la racine plutôt qu'une autre. En revanche, il peut être intéressant de coupler ce signal à d'autres signaux dépendants également de la position de la racine. Ainsi, les modèles de réconciliation d'arbres de gènes et d'espèces sont capables d'estimer la position de la racine de l'arbre des espèces qui minimise le nombre de transferts horizontaux de gènes (Abby et al., 2012) ou maximisent l'ensemble des probabilités des scénarios de réconciliations inférés sur un ensemble de gènes (Szöllősi et al., 2012). L'utilisation conjointe des deux signaux pourrait alors permettre de proposer efficacement une position de la racine d'un arbre d'espèces, qu'il serait intéressant de comparer avec la position proposée par la méthode classique du groupe externe.

Le champ de recherche ayant attrait à la résurrection de protéines ancestrales devrait, dans un futur proche, bénéficier des progrès méthodologiques présentés dans cette thèse pour estimer de meilleures séquences, mais également bénéficier de l'amélioration de la biologie synthétique, permettant de produire des milliers de protéines pour un coût humain et financier minimum. Si l'on s'intéresse à un gène en particulier pour lequel on cherche à retracer l'histoire évolutive en termes de structure ou de fonction, il sera possible de synthétiser en masse des protéines dont les séquences sont tirées des distributions postérieures des reconstructions, afin d'éviter de ne considérer que la séquence ML à chaque noeud. L'impact de l'incertitude de la reconstruction *in silico* sur les inférences fonctionnelles pourrait alors être mesuré afin d'affiner les conclusions biologiques. En outre, des centaines voire des milliers de protéines inférées comme étant

présentes chez un ou des ancêtres donnés le long d' un arbre d'espèces pourraient permettre d'estimer les intéractomes protéiques ancestraux à l'aide d'expériences de co-immunoprécipitation. Comprendre la dynamique évolutive de ces réseaux d'intéractions pourraient enfin permettre d'associer à ces réseaux des notions fonctionnelles et évolutives afin de mieux appréhender l'évolution des organismes en relation avec leur environnement. Enfin, au delà de la résurrection de protéines uniques (ou considérées comme étant indépendantes dans les perspectives mentionnées juste au dessus), il y a un fort intérêt à ressusciter des protéines en intéraction physique (au sein de complexes) ou fonctionnelle (au sein d'un pathway), afin de résurrecter des complexes ou des pathways protéiques. Seulement, la reconstruction des séquences ancestrales associées à ces protéines devra se faire à l'aide de gène co-réconciliés, prenant en compte à la fois les informations de duplication, transferts, pertes mais aussi les informations de co-évolution entre gènes/protéines.

# 6 Annexes

## 6.1 Nouvelle version des librairies Bio++.

### 6.1.1 Introduction

La majorité des travaux en Maximum de Vraisemblance que j'ai réalisé durant cette thèse ont été effectué avec les programmes appartenant à la suite de programmes bppSuite (Dutheil and Boussau, 2008), dépendant des librairies Bio++ (Dutheil et al., 2006). Ces programmes permettent de réaliser un grand nombre d'expériences phlogénétiques, allant de l'estimation de paramètres évolutifs le long d'un arbre avec bppML, à l'inférence de séquences ancestrales avec bppAncestor, en passant par la simulation de séquences avec bppSeqGen. Au delà de la phylogénie et de l'analyse d'arbres, les librairies Bio++ mettent à disposition toute une série de fonctionnalités dédiées à l'analyse de séquences biologiques, à la génétique des populations, à la requête de séquences dans les banques de données etc.

Dans l'article suivant, publié dans le journal Molecular Biology & Evolution, une nouvelle version des librairies est présentée, avec de nouvelles fonctionnalités par rapport à celles publiées en 2006. Ma participation a consisté en l'implémentation dans les librairies de plusieurs fonctions de routine, ainsi que l'implémentation du modèle COaLA (Groussin et al., 2013a) et

de modèles de mélanges (Le et al., 2008a; Le and Gascuel, 2010; Groussin et al., 2013b).

### 6.1.2 Manuscrit

MBE Advance Access published June 19, 2013

# Bio++: Efficient Extensible Libraries and Tools for Computational Molecular Evolution

Laurent Guéguen,[1] Sylvain Gaillard,[2,3,4] Bastien Boussau,[1,5] Manolo Gouy,[1] Mathieu Groussin,[1] Nicolas C. Rochette,[1] Thomas Bigot,[1] David Fournier,[6] Fanny Pouyet,[1] Vincent Cahais,[7] Aurélien Bernard,[7] Céline Scornavacca,[7] Benoît Nabholz,[7] Annabelle Haudry,[1] Loïc Dachary,[8] Nicolas Galtier,[7] Khalid Belkhir,[7] and Julien Y. Dutheil*[,7,9]

[1]Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, CNRS, INRIA, Villeurbanne, France

[2]INRA, Institut de Recherche en Horticulture et Semences, Angers, France

[3]Agrocampus Ouest, Institut de Recherche en Horticulture et Semences, Angers, France

[4]Institut de Recherche en Horticulture et Semences, Université d'Angers, LUNAM Université, Angers, France

[5]Department of Integrative Biology, University of California, Berkeley

[6]Computational Biology and Data Mining Group, Max-Delbrueck-Center for Molecular Medicine, Berlin, Germany

[7]Institut des Sciences de l'Evolution, Université Montpellier 2, Montpellier, France

[8]12 bd Magenta, Paris, France

[9]Department of Organismic Interactions, Max Planck Institute for Terrestrial Microbiology, Marburg, Germany

***Corresponding author:** E-mail: julien.dutheil@univ-montp2.fr.

**Associate editor:** Sudhir Kumar

## Abstract

**Efficient algorithms and programs for the analysis of the ever-growing amount of biological sequence data are strongly needed in the genomics era. The pace at which new data and methodologies are generated calls for the use of pre-existing, optimized—yet extensible—code, typically distributed as libraries or packages. This motivated the Bio++ project, aiming at developing a set of C++ libraries for sequence analysis, phylogenetics, population genetics, and molecular evolution. The main attractiveness of Bio++ is the extensibility and reusability of its components through its object-oriented design, without compromising the computer-efficiency of the underlying methods. We present here the second major release of the libraries, which provides an extended set of classes and methods. These extensions notably provide built-in access to sequence databases and new data structures for handling and manipulating sequences from the omics era, such as multiple genome alignments and sequencing reads libraries. More complex models of sequence evolution, such as mixture models and generic *n*-tuples alphabets, are also included.**

***Key words:*** **bioinformatics, models of sequence evolution, phylogeny, C++ libraries.**

The field of molecular evolution has always relied heavily on the use of computers for modeling and analysis (Eck and Dayhoff 1966; Fitch and Margoliash 1967). The need to use computers, and to use them efficiently, is even more pressing now that genome sequence data are accumulating at an increasing pace. In 2006, version 1.0.0 of the Bio++ libraries was published (Dutheil et al. 2006) with the aim to provide a set of flexible, efficient, object-oriented C++ methods for sequence analysis, population genetics, and molecular phylogenetics. Bio++ offers a set of ready-to-use bricks to construct sequence analysis pipelines, develop new complex probabilistic models, run maximum likelihood inference or simulate data, among other possibilities. Since their initial release the libraries have been used in a variety of published works and have enabled the development of new models and tools (for recent examples, see Bérard and Guéguen 2012; Caffrey et al. 2012; Dutheil et al. 2012; Szöllosi et al. 2012; Boussau et al. 2013; Groussin et al. 2013; Scornavacca et al. 2013). As they have been attracting new users and developers, the libraries have been extended to include new analysis tools, and now contain the largest set of models for sequence evolution ever implemented.

## New Developments

The initial release of the Bio++ libraries (Dutheil et al. 2006) was followed by several regular updates, and a major new version (Bio++ 2.0.0) was released in 2011. As of January 2013, the current stable version is 2.1.0. Since version 1.0.0, the libraries have extensively developed and new libraries were added to the initial set. These libraries provide new functionalities, mainly dedicated to database access, graphics and graphical user interfaces (GUIs), as well as genomic analysis. The original libraries have also been extended to incorporate new models and analytical tools.

### Architecture of the Libraries

Since version 1.0.0, the amount of code in the libraries has more than doubled, reaching a total of more than 700 classes.







Letter

For ease of use, the code is split into several libraries, which can be installed and linked independently, depending on the user's specific needs. Version 2.1.0 contains eight libraries. The "bpp-core" library contains basal classes and interfaces necessary for the development of applications with Bio++. Three other libraries inherited from version 1.0.0 gather tools for sequence analysis (bpp-seq), phylogenetics (bpp-phyl), and population genetics (bpp-popgen). Finally, the following four new libraries were developed:

- bpp-raa, for Remote Acnuc Access, providing classes to query sequence databases
- bpp-seq-omics and bpp-phyl-omics, providing classes for (phylo)genomic analyses
- bpp-qt, providing graphical components based on the Qt library (Blanchette and Summerfield 2008).

Figure 1 shows the dependencies between these libraries. We now briefly describe the recent developments of the original Bio++ components, and the content of the new ones.

## Numerical Tools

The models available in Bio++ require numerical routines, which are coded in the core library. Since Bio++ 1.0.0, the collection of available algorithms has been extended (e.g., we added support for numerical derivatives, function reparametrization, and sampling procedures), and the efficiency of existing methods has been further improved. The library provides a fully object-oriented implementation of commonly used routines and algorithms for function minimization and derivation, or matrix calculus. In particular, the library offers a large set of object-oriented, event-driven minimization algorithms for finely tuned optimization of complex functions with numerous parameters, such as likelihood under phylogenetic models. Developing new probabilistic models is now made easier thanks to a larger array of continuous or discretized distributions (Gaussian, exponential, beta, Dirichlet, and any mixture of distributions), as well as standard algorithms for hidden Markov modeling (forward, backward algorithms, and posterior decoding with rescaling to avoid numerical underflow [Durbin et al. 1998]).

## Database Access

The bpp-raa library allows network access to several nucleotide and protein sequence databases, both generalist ones (the EMBL sequence library, GenBank, and UniProt) and databases of families of homologous protein-coding genes (e.g., HOGENOM, HOMOLENS, HOVERGEN; Penel et al. 2009). Bpp-raa employs the ACNUC sequence retrieval system (Gouy and Delmotte 2008) to communicate between the library user and a sequence database. The `bpp::RAA` class opens a network connection to a given database, and allows extracting sequences and annotations based on sequence name or accession number, with optional translation to protein. This class also allows building the list of sequences that match a given query, including complex queries that involve logical combinations of criteria (e.g., species name AND/OR/NOT keyword AND/OR/NOT reference AND/OR/NOT previous query). The members of a sequence list can then be extracted for local processing. The `bpp::RaaSpeciesTree` class allows using the taxonomy associated with sequence databases, walking up and down this tree, and finding its nodes by name or numerical taxon ID.

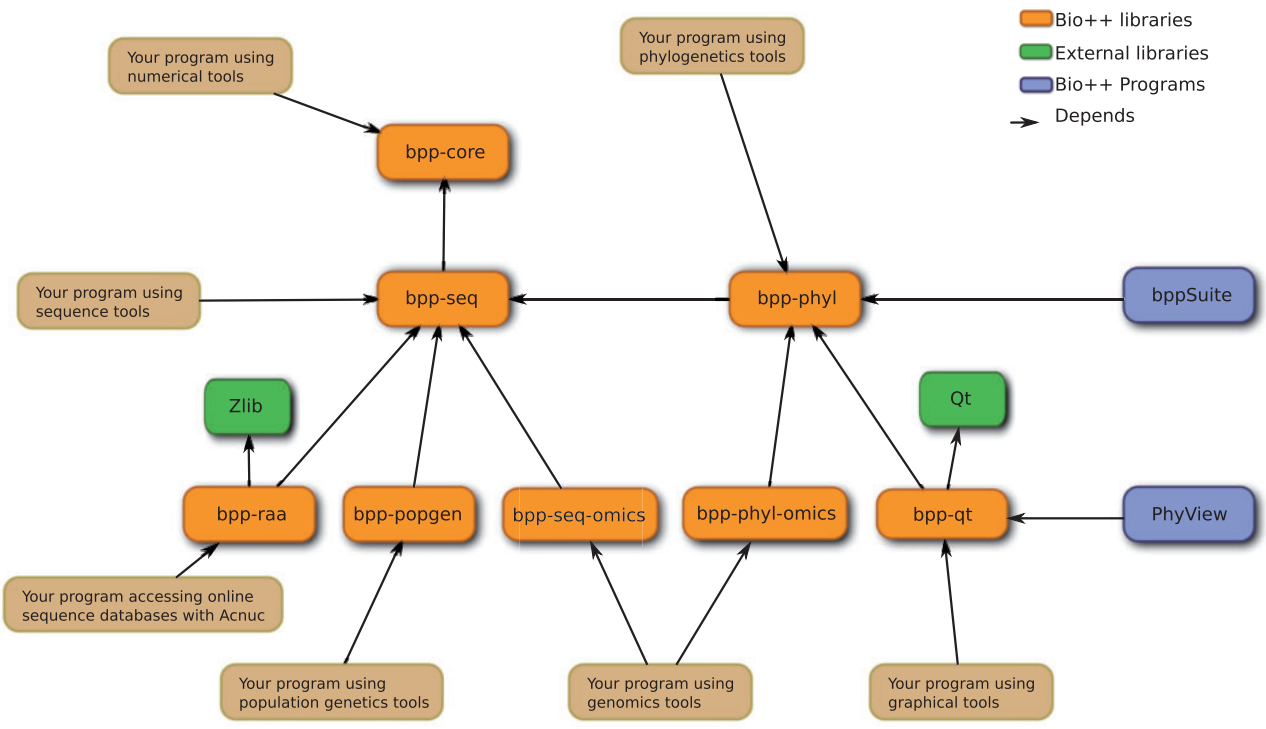



FIG. 1. Dependencies between libraries and programs.

## Genomic Tools

The sequence class hierarchy has been extended to cope with the increasing amount of genomic data. These developments follow three main axes: 1) faster handling of sequences, notably via the use of binary coding to allow more efficient comparisons, and rewriting of file parsers, 2) support for subsequences and features, including parsers for GFF and GTF formats, as well as storage and manipulation of meta-data like quality scores, and 3) addition of new file formats, notably those used for (Next Generation) sequencing (Phred, FastQ, and MAF). These new data structures enable a very efficient parsing and filtering of typical genomic data sets. A simple program using a `bpp::SequenceIterator` based on the new `bpp::FastQ` parser and the `bpp::SequenceWithQuality` data structure is able to parse 20 millions paired-ends reads of 100 bp in 20 min on a desktop computer, whereas the same analysis requires more than 1 h and 30 min with an equivalent pipeline built using the (locally installed) Galaxy platform (Hillman-Jackson et al. 2012).

## New Models of Sequence Evolution

The first version of Bio++ already supported a large variety of models of sequence evolution for nucleotide or amino acid sequences, later extended with branch-heterogeneous models (Dutheil and Boussau 2008; Groussin et al. 2013). Version 2.1.0 offers in addition a generalized modeling framework that considers *n*-tuples as the evolving units in a sequence. This permits an extensible and flexible implementation of codon models. Specific features of codon models are implemented separately in abstract classes, enabling the development of customized codon models. The currently implemented models notably support 1) position-specific substitution rates; 2) biochemical distances between the encoded amino acids (as in the GY94 model; Goldman and Yang 1994); and 3) preferences between synonymous codons (Yang and Nielsen 2008). Equilibrium frequencies are modeled either in a position-specific manner or at the codon level, with possibility for the user to provide his/her own implementation. The substitution rate is proportional either to the target codon equilibrium frequencies (as in GY94) or to the target nucleotide equilibrium frequencies (as in the MG94 model; Muse and Gaut 1994). This generic implementation unifies the vast majority of models proposed in the literature (Pond and Muse 2005; Wong et al. 2006; Mayrose et al. 2007).

Bio++ 2.0.0 also provides support for mixed models. In these models, a site can "choose" between several models (Yang and Wang 1995; Yang et al. 2000). The resulting compound likelihood for a site is the average of the conditional likelihoods for each model, weighted by their probability distribution. Using this new generic framework, several previously published mixed models have been made available in Bio++, such as codon models M1, M2, M3, M7, and M8 from the widely used codeml program (Yang et al. 2000; Yang 2007) for modeling site-specific selection coefficients or the protein models UL2, UL3, EX2, CAT-C10 to C60 among others (Le, Lartillot, et al. 2008; Le, Gascuel, et al. 2008) for modeling site-specific properties of proteins.

A generic framework has also been implemented for combining branch-heterogeneous models with mixed models. In this framework, it is possible to assign mixed models to a subset of branches. Different mixed models can be assigned to separate branches, in which case a site is allowed to switch between categories of models at nodes, as in the branch-site model of PAML (Zhang et al. 2005) (fig. 2 and supplementary fig. S1, Supplementary Material online). In addition, it is possible to constrain those switches so that particular sets of branches are always in the same category. The current implementation therefore covers a large set of mixed models available in the literature, whilst enabling the development of new ones.

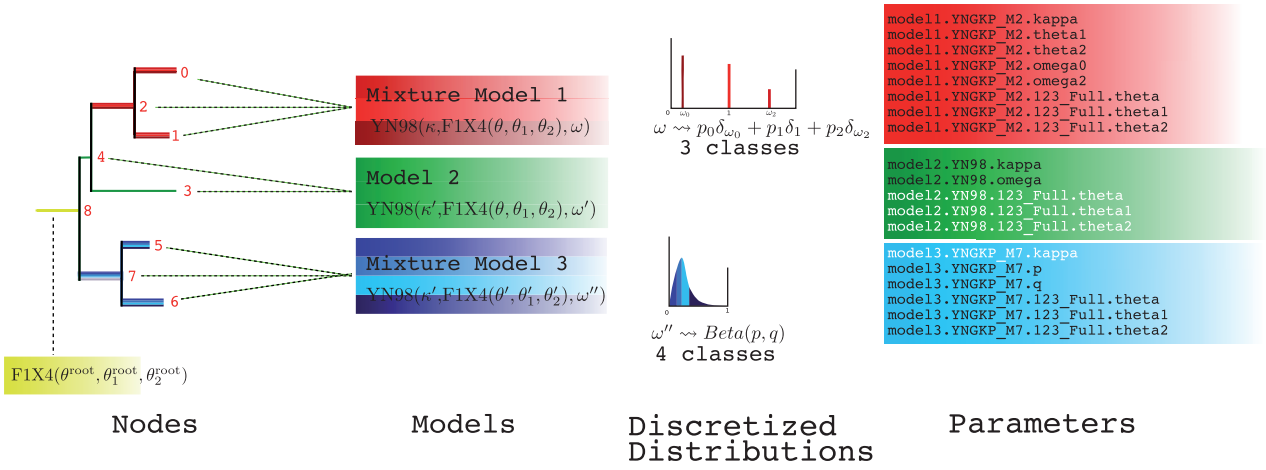



**Fig. 2.** Non-homogeneous modeling with mixture models. Example of nonstationary and nonhomogeneous modeling of evolution of a codon sequence, using three models (M0, M2, and M7) as defined in Nielsen and Yang (1998) and Yang et al. (2000). On branches 0, 1, and 2, a site can choose between three YN98 models, in which omega can be <1, =1, or >1, with specific probabilities. On branches 5, 6, and 7, a site can choose between four YN98 models, in which omega follows a discretized beta distribution. The equilibrium frequencies of the model on branches 3 and 4 are the same as the ones of the model on branches 0, 1, and 2. The kappa parameter value is the same on branches 3, 4, 5, 6, and 7. In the parameter list, parameters in white are shared between models. Although artificial, this example demonstrates the generality of the modeling framework implemented in Bio++.

### An Extended Set of Tools for Molecular Evolution

The large set of models of sequence evolution available in Bio++ can be used in combination with routinely used methods in evolutionary bioinformatics, such as tree-building, population analyses, sequence simulation, and ancestral state reconstruction. Although the libraries are not dedicated to phylogenetic reconstruction per se (for which specialized software exist), they contain building blocks based on published algorithms which can be useful to develop new methods in that field. Such "blocks" include parsimony score and tree likelihood computation (with simple and double recursive algorithms, see Felsenstein 2003), as well as nearest-neighbor interchange topology movements. Distance methods are also available, including neighbor joining (Saitou and Nei 1987) and BioNJ (Gascuel 1997), which are implemented in an object-oriented way. The majority of models implemented can also be used to simulate sequences, including covarion models and nonhomogeneous models, and to reconstruct ancestral sequences using the empirical Bayesian approach (Yang et al. 1995). Population genetics statistics include the computation of a variety of sequence diversity estimators, Tajima's D (Tajima 1989), neutrality index (Rand and Kann 1996) and McDonald and Kreitman's count table for testing of positive selection (McDonald and Kreitman 1991). Since version 1.0.0, a notable addition is the development of generic substitution mapping procedures (Minin and Suchard 2008; Tataru and Hobolth 2011), which can be used to characterize patterns of substitution in a robust and efficient manner (Lemey et al. 2012; Romiguier et al. 2012).

### Graphical Tools

Graphical tools have been introduced in version 2.0.0 of the libraries. The bpp-core library provides a generic `bpp::GraphicDevice` class supporting drawing operations such as lines, polygons, and text writing, as well as dedicated interfaces to handle colors and fonts. The bpp-core library includes three implementations of this interface: Scalable Vector Format, LaTeX's Portable graphic Format and the Xfig format, and the bpp-qt library provides an additional implementation based on the Qt graphic library. The bpp-phyl library contains several algorithms for plotting trees on a `bpp::GraphicDevice`, which can therefore be used to save a graphical representation of a tree into a file, or as part of a GUI. Pre-built GUI components for phylogenetic tree browsing are included in the bpp-qt library, and used in the bppPhyView software, a powerful Bio++ based tree editor.

## The Bio++ Program Suite and the BppO Language

Several programs developed using the Bio++ libraries are distributed as the Bio++ Program Suite (bppSuite), including the following:

- bppML, which performs maximum likelihood estimation of models of sequence evolution,
- bppSeq, which simulates sequences under a model of sequence evolution,
- bppAncestor, which reconstructs ancestral sequences,
- bppDist, which reconstructs phylogenies based on distance matrices.

They all share a common language for the description of their parameters, notably models of sequence evolution. In Bio++ 2.1.0, this language has a dedicated Application Programming Interface (API) included in the library. It is referred to as the Bio++ Options language, or simply BppO. With BppO, one can easily specify which of the input/output formats, models, frequencies, discrete distributions, to use and perform—depending on the chosen bppSuite program—maximum likelihood estimation of parameters, ancestral sequence reconstruction, sequence simulation, and so forth. Two examples showing how complex models can be specified using the BppO syntax are given in figures 2 and 3 (for codon models) and supplementary figures S1 and S2, Supplementary Material online (for nucleotide models). Programs in bppSuite output their results in a BppO file, which can then be used directly as input for another program. This makes it easy for instance to use a previously fitted model to simulate sequences or reconstruct ancestral sequences. Through the BppO language and BppSuite, a large set of the features of Bio++ are made available to the user without the need for C++ programming.

## Availability and Future Directions

The Bio++ libraries are distributed under the CeCILL 2.0 license (compatible with the GNU Public License) at http://bioweb.me/biopp (last accessed June 6, 2013). Source code can be compiled (at least) on any system where the GNU compiler collection is available (including Linux, MacOS, and Windows). Bio++ uses CMake for its configuration (Martin and Hoffman 2010), which facilitates its integration with widely used development environments such as Visual Studio, XCode, CodeBlocks, or Eclipse. Stable versions are released yearly, with precompiled and source packages available for the most common Linux distributions and MacOS. Since 2011, the Bio++ libraries and packages are also directly available from the Debian distribution (and therefore its derivatives such as Ubuntu and Linux Mint). The latest development version of the code can be obtained from a central Git repository.

Bio++ uses unit tests and is checked nightly. The API documentation, generated using the Doxygen program (http://www.doxygen.org, last accessed June 6, 2013), is also updated nightly and made available online to ease the development of new applications. In addition, the Bio++ website features a wiki-based documentation, example programs, a bug tracker and two forums (biopp-help dedicated to getting help with the use of the libraries, with more than 70 members, 160 topics, 890 posts, and biopp-devel for general development discussion, with more than 30 members, 260 topics, 940 posts).

Thanks to its growing community, Bio++ is under continuous development. The strength of Bio++ is its combination of generality and efficiency. Generality is achieved through the

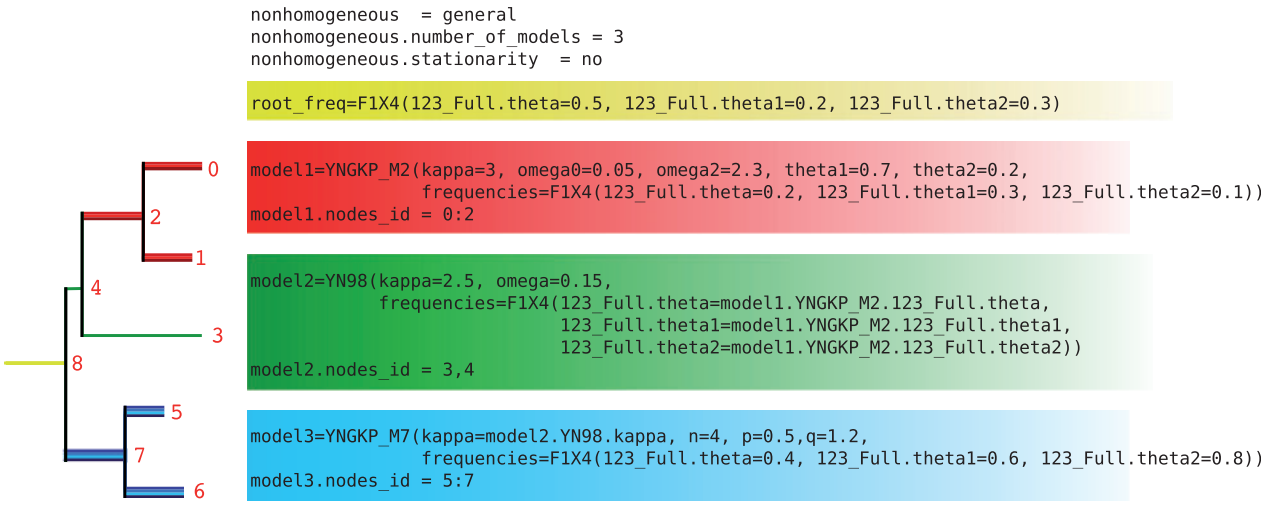



**Fig. 3.** Syntax of the modeling in the BppO language, using the specific names of models described in the bibliography.

strictly object-oriented design of the library, which eases the development of new models of sequence evolution.

Comparison with other pieces of software shows that the versatility of Bio++ comes at a minimal cost in terms of computer resources (supplementary table S1, Supplementary Material online). For instance, on a nucleotide data set of 79 sequences and 2,353 sites, the BppML program (from the Bio++ program suite) fits a GTR substitution model with 4 gamma-distributed rate classes in 2′05 minutes on a linux desktop machine, using 215 kB of memory. PhyML achieves the same analysis in 1 minute 18 seconds with 390 kB. PAML uses only 28 kB but performs the estimation in 6 minutes 54 seconds. All three programs return the same parameter estimates and likelihood. This efficiency is due to a fine control of memory usage achieved through the classes and tools of the C++ Standard Template Library, as well as the efficient function optimizers implemented in bpp-core. For phylogenetic models, a dedicated modified Newton–Raphson algorithm is used, based on an initial idea from Felsenstein's phylip package, further improved in the NHML software, and re-implemented in an object-oriented manner in Bio++. Programs developed with Bio++ are therefore well fitted for data analyses typically achieved by their C-coded, non-library-based counterparts. Further improving the performances of the libraries is one of the next challenges that the Bio++ developers are currently pursuing, notably by pushing the limit of numerical underflow and developing support for parallelization, to handle increasingly large data sets.

## Supplementary Material

Supplementary figures S1 and S2 and table S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

## Acknowledgments

The authors thank the users of the Bio++ forums for their continuous feedback helping them to improve the code of the library, as well as the reviewers for their constructive comments on an earlier version of this manuscript. This work has been partially funded by the Agence Nationale de la Recherche ANCESTROME project (ANR-10-BINF-01-01 and ANR-10-BINF-01-02) and the European Research Council PopPhyl project (ERC 232971). This publication is the contribution no. 2013-059 of the Institut des Sciences de l'Evolution de Montpellier (ISE-M).

## References

Bérard J, Guéguen L. 2012. Accurate estimation of substitution rates with neighbor-dependent models in a phylogenetic context. *Syst Biol.* 61:510–521.

Blanchette J, Summerfield M. 2008. C++ GUI programming with Qt 4, 2nd ed. Upper Saddle River (NJ): Prentice Hall.

Boussau B, Szöllosi GJ, Duret L, Gouy M, Tannier E, Daubin V. 2013. Genome-scale coestimation of species and gene trees. *Genome Res.* 23:323–330.

Caffrey BE, Williams TA, Jiang X, Toft C, Hokamp K, Fares MA. 2012. Proteome-wide analysis of functional divergence in bacteria: exploring a host of ecological adaptations. *PLoS One* 7:e35659.

Durbin R, Eddy SR, Krogh A, Mitchison G. 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge (UK): Cambridge University Press.

Dutheil J, Boussau B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol Biol.* 8:255.

Dutheil J, Gaillard S, Bazin E, Glémin S, Ranwez V, Galtier N, Belkhir K. 2006. Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinformatics* 7:188.

Dutheil JY, Galtier N, Romiguier J, Douzery EJP, Ranwez V, Boussau B. 2012. Efficient selection of branch-specific models of sequence evolution. *Mol Biol Evol.* 29:1861–1874.

Eck RV, Dayhoff MO. 1966. Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. *Science* 152:363–366.

Felsenstein J. 2003. Inferring phylogenies, 2nd ed. Sunderland (MA): Sinauer Associates.

Fitch WM, Margoliash E. 1967. Construction of phylogenetic trees. *Science* 155:279–284.

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol.* 14: 685–695.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol.* 11: 725–736.
Gouy M, Delmotte S. 2008. Remote access to ACNUC nucleotide and protein sequence databases at PBIL. *Biochimie* 90:555–562.
Groussin M, Boussau B, Gouy M. Forthcoming 2013. A branch-heterogeneous model of protein evolution for efficient inference of ancestral sequences. *Syst Biol.*
Hillman-Jackson J, Clements D, Blankenberg D, Taylor J, Nekrutenko A. 2012. Using galaxy to perform large-scale interactive data analyses. *Curr Protoc Bioinformatics.* Chapter 10:Unit10.5.
Le SQ, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24:2317–2323.
Le SQ, Lartillot N, Gascuel O. 2008. Phylogenetic mixture models for proteins. *Philos Trans R Soc Lond B Biol Sci.* 363:3965–3976.
Lemey P, Minin VN, Bielejec F, Kosakovsky Pond SL, Suchard MA. 2012. A counting renaissance: combining stochastic mapping and empirical Bayes to quickly detect amino acid sites under positive selection. *Bioinformatics* 28:3248–3256.
Martin K, Hoffman B. 2010. Mastering CMake: a cross-platform build system Version 5. Villeurbanne (France): Kitware.
Mayrose I, Doron-Faigenboim A, Bacharach E, Pupko T. 2007. Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates. *Bioinformatics* 23: i319–i327.
McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351:652–654.
Minin VN, Suchard MA. 2008. Fast, accurate and simulation-free stochastic mapping. *Philos Trans R Soc Lond B Biol Sci.* 363: 3985–3995.
Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol.* 11: 715–724.
Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148:929–936.
Penel S, Arigon A-M, Dufayard J-F, Sertier A-S, Daubin V, Duret L, Gouy M, Perrière G. 2009. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics* 10(6 Suppl):S3.
Pond SK, Muse SV. 2005. Site-to-site variation of synonymous substitution rates. *Mol Biol Evol.* 22:2375–2385.
Rand DM, Kann LM. 1996. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. *Mol Biol Evol.* 13:735–748.
Romiguier J, Figuet E, Galtier N, Douzery EJP, Boussau B, Dutheil JY, Ranwez V. 2012. Fast and robust characterization of time-heterogeneous sequence evolutionary processes using substitution mapping. *PLoS One* 7:e33852.
Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 4:406–425.
Scornavacca C, Paprotny W, Berry V, Ranwez V. 2013. Representing a set of reconciliations in a compact way. *J Bioinform Comput Biol.* 11:1250025.
Szöllosi GJ, Boussau B, Abby SS, Tannier E, Daubin V. 2012. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc Natl Acad Sci U S A.* 109: 17513–17518.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.
Tataru P, Hobolth A. 2011. Comparison of methods for calculating conditional expectations of sufficient statistics for continuous time Markov chains. *BMC Bioinformatics* 12:465.
Wong WSW, Sainudiin R, Nielsen R. 2006. Identification of physico-chemical selective pressure on protein encoding nucleotide sequences. *BMC Bioinformatics* 7:148.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.
Yang Z, Kumar S, Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141:1641–1650.
Yang Z, Nielsen R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol.* 25:568–579.
Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449.
Yang Z, Wang T. 1995. Mixed model analysis of DNA sequence evolution. *Biometrics* 51:552–561.
Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol.* 22:2472–2479.

## 6.2 La datation des temps de divergence des Foraminifères benthiques.

### 6.2.1 Introduction

Depuis une quinzaine d'années, un nombre florissant de publications abordent l'estimation des âges de divergences à partir des données moléculaires, en utilisant des modèles permettant à la fois une variation des taux d'évolution dans le temps et la prise en compte de multiples calibrations fossiles (Thorne et al., 1998; Drummond et al., 2006; Yang and Rannala, 2006; Heath, 2012). Ces nouveaux modèles ont permis d'améliorer les inférences de datation par rapport à l'approche de datation par horloge moléculaire stricte (stipulant une constance des taux dans le temps) à l'aide d'un point de calibration unique, qui fut critiquée avec virulence à la fin des années 1990, début des années 2000 (Hedges et al., 1996; Bromham et al., 1998; Graur and Martin, 2004) pour largement sur-estimer les dates de divergence. Afin de mieux comprendre ces modèles et la façon de réaliser une datation moléculaire à l'aide de données fossiles, j'ai effectué un stage lors de mon Master dans le laboratoire de Ziheng Yang à Londres. Avec Brune Rannala, ils ont publié en 2006 et 2007 le programme Mcmctree et un modèle d'horloge relachée prenant en compte l'incertitude sur les calibrations fossiles (Yang and Rannala, 2006; Rannala and Yang, 2007). J'ai utilisé cette approche de datation pour comprendre la dynamique de spéciation des Foraminifères benthiques. Les Foraminifères benthiques sont doublement intéressants car ils sont l'un des groupes d'eukaryotes unicellulaires les plus abondants sur Terre mais aussi parce qu'il est très difficile de les dater du fait de leur évolution très rapide au niveau moléculaire. J'ai établi une collaboration avec jan Pawlowski, spécialiste de l'évolution et de la phylogénie des Foraminifères et des Eukaryotes unicellulaires en général.

L'article suivant, publié dans Molecular Phylogenetics & Evolution, a permis de proposer une date d'émergence du clade des Foraminifères benthiques, pré-datant de quelques 200 millions d'années l'apparition du premier spécimen de Foraminifères dans le registre fossile. L'étude a également permis de dater la divergence de clades majeurs, tels que les Rotaliids et les Miliolids.

### 6.2.2 Manuscrit

Molecular Phylogenetics and Evolution 61 (2011) 157–166

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

# Bayesian relaxed clock estimation of divergence times in foraminifera

Mathieu Groussin [a,1], Jan Pawlowski [b], Ziheng Yang [a,c,*]

[a] Galton Laboratory, Department of Biology, University College London, London WC1E 6BT, England, United Kingdom
[b] Department of Genetics and Evolution, University of Geneva, CH-1211 Geneva, Switzerland
[c] Centre for Computational and Evolutionary Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China

ARTICLE INFO

*Article history:*
Received 13 February 2011
Revised 17 May 2011
Accepted 10 June 2011
Available online 23 June 2011

*Keywords:*
Foraminifera
Divergence times
Bayesian method
Relaxed clock
MCMC
Infinite-sites plot

ABSTRACT

Accurate and precise estimation of divergence times during the Neo-Proterozoic is necessary to understand the speciation dynamic of early Eukaryotes. However such deep divergences are difficult to date, as the molecular clock is seriously violated. Recent improvements in Bayesian molecular dating techniques allow the relaxation of the molecular clock hypothesis as well as incorporation of multiple and flexible fossil calibrations. Divergence times can then be estimated even when the evolutionary rate varies among lineages and even when the fossil calibrations involve substantial uncertainties. In this paper, we used a Bayesian method to estimate divergence times in Foraminifera, a group of unicellular eukaryotes, known for their excellent fossil record but also for the high evolutionary rates of their genomes. Based on multigene data we reconstructed the phylogeny of Foraminifera and dated their origin and the major radiation events. Our estimates suggest that Foraminifera emerged during the Cryogenian (650–920 Ma, Neo-Proterozoic), with a mean time around 770 Ma, about 220 Myr before the first appearance of reliable foraminiferal fossils in sediments (545 Ma). Most dates are in agreement with the fossil record, but in general our results suggest earlier origins of foraminiferal orders. We found that the posterior time estimates were robust to specifications of the prior. Our results highlight inter-species variations of evolutionary rates in Foraminifera. Their effect was partially overcome by using the partitioned Bayesian analysis to accommodate rate heterogeneity among data partitions and using the relaxed molecular clock to account for changing evolutionary rates. However, more coding genes appear necessary to obtain more precise estimates of divergence times and to resolve the conflicts between fossil and molecular date estimates.



## 1. Introduction

Foraminifera belong to the eukaryotic super-group Rhizaria and are arguably one of the most important protist groups on Earth (Keeling et al., 2005). They are widely known for inhabiting marine ecosystems, but occupy freshwater and terrestrial environments as well (Holzmann et al., 2003; Lejzerowicz et al., 2010). In ocean habitats, foraminifers can have benthic or planktonic mode of life. The phylogenetic relationship between benthic and planktonic species are currently controversial (Ujiié et al., 2008) but planktonic species are known to appear later in the fossil record. A hallmark of foraminifers is their shells, more specifically termed "tests", which can either be organic, agglutinated or calcareous. Further variations in test morphology exist between taxa due to the construction of structurally distinct unilocular (single chamber) and multilocular (multi-chambers) tests. In general, the organic walls, which are present only in unilocular taxa, are thin and consist of an association between proteins and mucopolysaccharides. The agglutinated foraminifers form their test by cementing environmental particles and may have one or several chambers. Finally, calcareous foraminifers secrete calcium carbonate, principally as calcite, to constitute the wall of single- or multilocular tests.

Foraminifera possess one of the most profuse fossil records among eukaryotes. The earliest Cambrian foraminiferal genus *Platysolenites* has the appearance of a large, simple, agglutinated tube resembling modern foraminiferal genus *Bathysiphon* (McIlroy et al., 2001). Other straight and coiled tubular agglutinated foraminifera, including genus *Ammodiscus* have been reported from the Lower and Middle Cambrian (Culver, 1991). Some studies mention the possibility that unilocular agglutinated foraminifers were already present during the Upper Vendian (Ediacaran period), at the end of the Neo-Proterozoic (Gaucher and Sprechmann, 1999). However, this proposal is controversial given the difficulty in attributing these fossils to Foraminifera with confidence. A recent

* Corresponding author. Address: Department of Biology, University College London, Darwin Building, Gower Street, London WC1E 6BT, England, United Kingdom. Fax: +44 (20) 7679 7096.
*E-mail address:* z.yang@ucl.ac.uk (Z. Yang).
[1] Present address: Université de Lyon, F-69000, Lyon, Université Lyon 1, CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622 Villeurbanne, France.

1055-7903/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2011.06.008

study (Pawlowski et al., 2003) proposed that Foraminifera actually emerged during the Neo-Proterozoic, but was unable to provide a more concise time interval for the emergence than between 690 and 1150 Ma. However, this approach, based on molecular divergence time estimates, may have been limited by the exclusive use of partial ribosomal SSU sequences.

Until recently, molecular phylogeny of Foraminifera was inferred almost solely from ribosomal DNA sequences (Pawlowski et al., 1994, 1997, 1999, 2003; Ertan et al., 2004; Schweizer et al., 2008), with only a few phylogenetic analyses of protein coding genes (Habura et al., 2006; Longet and Pawlowski, 2007). For the majority of foraminiferal species, the only sequences available are the partial SSU rDNA sequences, characterized by variable substitution rates and unequal sequence lengths due to numerous insertion events (Pawlowski et al., 1997; Pawlowski and Holzmann, 2002). Consequently, the SSU rDNA alignments suffer from intense removal of sites, which leads to phylogenetic trees with weak resolutions. In spite of these drawbacks, the general view of foraminiferal phylogeny derived from these alignments is congruent in many respects with paleontological data. According to this view, the basal foraminiferal group is composed of monothalamiid (unilocular) organic-walled or agglutinated species, which gave rise to polythalamous (multilocular) clades at least twice during their history (Pawlowski and Holzmann, 2002). One of the resulting clades groups together agglutinated Textulariida and calcareous Rotaliida, while the other clade groups all lineages with early tubular chambers, including Miliolida, Spirillinida and some Lituolinida (*Ammodiscus*, *Miliammina*) (Pawlowski et al., 2003) (see Fig. 1). The earliest multichambered agglutinated Textulariida arose in the Devonian (>400 Ma), while the calcareous porcellaneous Miliolida are known since Carboniferous (>350 Ma) (Haynes, 1981). Nevertheless, the deep relationships between monothalamous and polythalamous lineages remain unsolved.

The molecular clock hypothesis provides a seductively powerful way to date evolutionary events such as speciation. However, the use of a strict molecular clock can lead to seriously biased estimates of divergence times when the clock is violated. The development of new algorithms in the likelihood and Bayesian frameworks has allowed different lineages to have variable evolutionary rates, thus improving the estimation of divergence times and reconciling palaeontologists and molecular systematists (Hasegawa et al., 2003; Douzery et al., 2004; Bell and Donoghue, 2005).

The well-studied fossil record of Foraminifera provides valuable information for calibrating the molecular phylogeny, to date the emergence and divergence times of major taxa. We should bear in mind, however, that the origin of Foraminifera is undoubtedly prior to their first appearance in the fossil record (545 Ma, Culver, 1991; McIlroy et al., 2001) and that this time difference is key in furthering our understanding of the early dynamics of Eukaryotes. To this end, the Bayesian statistical framework is of great interest because it permits the use of prior knowledge about times and rates. Thus, MCMCTREE in the PAML package (Yang and Rannala, 2006; Rannala and Yang, 2007; Yang, 2007) accommodates the uncertainties present in the fossil record by the use of soft bounds and flexible statistical distributions. In addition, MCMCTREE relaxes the molecular clock hypothesis by implementing two models of variable rates among lineages: the independent-rates model, where the rates for branches are independent variables from the same distribution, and the correlated-rates model, where the evolutionary rate of the daughter branch depends on the rate of the ancestral branch.

In this study, we used complete SSU rDNA sequences and three nuclear gene sequences, in addition to partial SSU rDNA sequences, to infer the phylogeny of Foraminifera and to estimate divergence times of their major lineages. We focused on benthic foraminifers because of the extreme evolutionary distances within planktonic species and between benthic and planktonic species. We conducted an extensive robustness analysis to examine the impact of various prior assumptions on our posterior time estimates. We discuss the conflicts between the molecular and fossil time estimates.

## 2. Material and methods

### 2.1. Sequence data

The main data set consists of 34 species representing the known taxonomic diversity of Foraminifera. Whenever possible, complete SSU sequences were extracted from the GenBank. When these are unavailable, partial SSU sequences were used. Three nuclear protein-coding genes were also included: actin-2, β-tubulin and RPB1. The actin-2 and β-tubulin sequences of *Quinqueloculina sp.* were retrieved by a blast procedure among all Expressed Sequence Tags (ESTs) available for this species. To reduce the proportion of missing data in the coding-genes data set, some species belonging to the same genus were merged and referred to by their genus names (*Ammonia*, *Allogromia*, *Bathysiphon*, *Bolivina*, *Reophax* and *Sorites*) (Suppl. Table 1). The SSU rDNA sequence is available for 34 species, while at least one coding-gene sequence is available for only 26 species. See Suppl. Table 1 for details.

### 2.2. Alignment

All sequences were aligned with Muscle 3.7 (Edgar, 2004). The SSU alignment was then improved by successive re-alignments with Muscle within regions of large insertions located between parts of conserved regions, by using the latest version of SeaView (Gouy et al., 2010). Instead of manually improving the SSU alignment by removing fast-evolving regions, we used the SlowFaster program (Kostka et al., 2008) to do so automatically. This requires a prior topology and uses maximum parsimony to establish different thresholds of evolutionary rates within pre-defined monophyletic groups. Alignments of different sizes were generated, depending on the number of fast sites removed, and for each alignment, the maximum likelihood (ML) tree was reconstructed using PhyML (Guindon and Gascuel, 2003) under the GTR + $\Gamma_5$ + I model with 100 bootstrap replicates. We chose for all later analysis of this paper the alignment that showed the highest bootstrap support values for the following nodes: origins of rotaliids and recent miliolids (*Quinqueloculina sp.*, *Pyrgo peruviana*, *Peneroplis pertusus*, *Marginopora vertebralis*, *Sorites* and *Amphisorus hemprichii*) and the origin of the group *Ammonia* + *Haynesina germanica* + *Elphidium williamsoni*. Those nodes were chosen because they were widely accepted to be monophyletic. This heuristic approach was taken mainly because numerous insertions and fast-evolving sites cause difficulties in the alignment. As the topology for the entire foraminiferal domain contains uncertainties, the input topology used by SlowFaster included multifurcations at unresolved nodes and respected the well-known and supported monophyly of rotaliids and recent miliolids. The Slow–Fast approach was not used to obtain the coding-genes alignment, as it is straightforward to obtain reliable alignments.

As a result, several data sets were compiled. SSU rDNA with 1942 sites, and the first and second codon positions of the three coding genes with 2148 sites. For phylogeny reconstruction, the data are analyzed using two partitioning strategies. In the first, the SSU rDNA data (1942 sites) and the first and second codon positions data (2148 sites) were concatenated into one partition, with 4099 sites in total. The second strategy treats the data as two partitions: the SSU rDNA vs. positions 1 + 2 of the coding genes. The dating analysis was conducted using the MCMCTREE

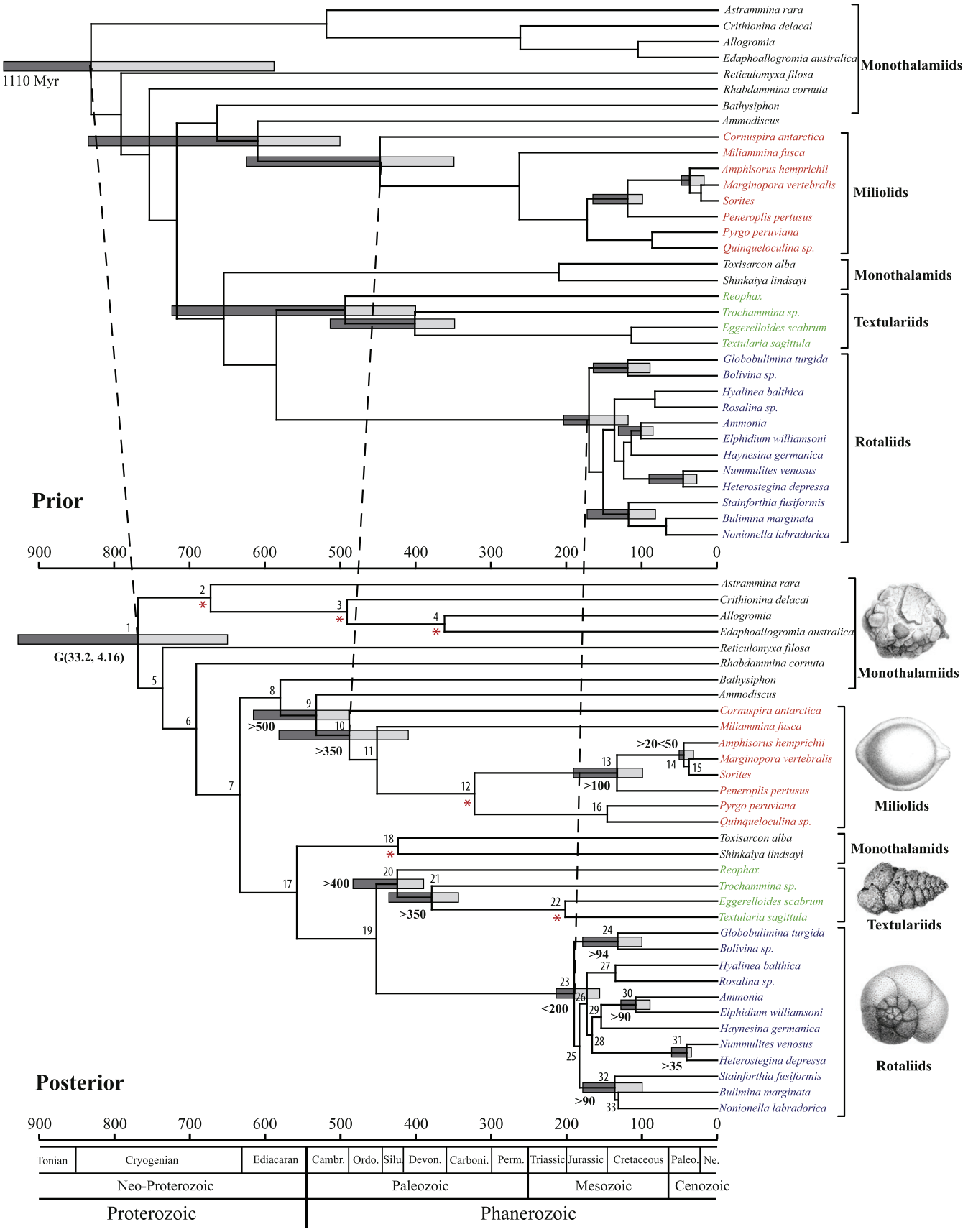



**Fig. 1.** The phylogeny of Foraminifera showing fossil calibrations and prior (top) and posterior (bottom) means of divergence times. Eleven fossil calibrations are used in the dating analysis, including nine minimum bounds, a maximum bound of 200 Myr for the root of Rotaliids and a pair of joint bounds (>20 < 50) for the origin of Soritinae. A gamma prior *G*(33.2, 4.16) is assigned on the age of the root. The 95% credibility intervals are shown for all nodes where fossil calibrations are available.

program, with the data analyzed as two partitions (rDNA vs. positions 1 + 2 of the coding genes) or four partitions (rDNA vs. three partitions for the three genes). Each partition had its own set of substitution parameters and substitution rates (Yang, 1996). Divergence times were also estimated using two mixed partitions (rDNA vs. amino acids).

### 2.3. Phylogenetic reconstructions

For the concatenated data set, the ML tree was reconstructed using the PhyML program (Guindon and Gascuel, 2003) under the GTR + $\Gamma_5$ + I model, with 100 bootstrap replicates. The Bayesian tree was reconstructed using PhyloBayes (Lartillot et al., 2009) under the GTR + CAT + $\Gamma_5$ model. Trace plots of the log-likelihood values confirmed convergence of the MCMC. Two MCMC runs were used to confirm consistency between runs. We also conducted a phylogenetic analysis on the two-partitions data set (see above) with RAxML (Stamatakis, 2006) and MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). The GTR + $\Gamma_5$ model, with 100 bootstrap replicates was used in RAxML, starting from a random tree. Similarly, the GTR + $\Gamma_5$ model was applied with MrBayes.

### 2.4. Calibration points and Bayesian divergence time estimation

Both the ML and Bayesian analyses produced unrooted trees. These were integrated with previous analyses of molecular and fossil data to generate a rooted tree for Bayesian estimation of divergence times using MCMCTREE.

Twelve fossil calibrations were used to calibrate the foraminiferal tree, implemented using soft bounds (Yang and Rannala, 2006) (Table 1). Nine of them represent minimum bounds and are assigned the truncated Cauchy distribution with $p = 0.1$ and $c = 0.2$ so that the density falls off rapidly away from the mode, which is close to the fossil minimum (Inoue et al., 2010). The microfossil record is well documented, so we expect the true age to be close to this minimum. The fossil calibrations are according to Haynes (1981) and Loeblich and Tappan (1987). A few comments are in order. (i) According to the fossil record, the radiation of Rotaliida occurred in Early Cretaceous but they probably already diverged during the Jurassic period (Haynes, 1981). Thus a maximum bound of 200 Ma was used on this node (node 23 in the tree of Fig. 1). (ii) The first presence of the Soritinae subfamily (*A. hemprichii*, *Marginopora* and *Sorites*) is recorded in rocks since 20 Ma, during Miocene (Haynes, 1981). On this node (node 14), a maximum bound of 50 Ma was applied, based on the fossil apparition of the earliest Archaiasinae, sister group of Soritinae (Holzmann et al., 2001). (iii) The program MCMCTREE requires a constraint on the age of the root in the dating analysis. As no reliable fossil calibration exists for the origin of Foraminifera, a gamma prior was used for the root age. A previous study based on molecular divergence time estimations proposed that Foraminifera could have originated between 690 and 1150 Ma (Pawlowski et al., 2003). The oldest fossils attributed to Foraminifera with confidence are from the early Cambrian (Culver, 1991). So we used the gamma prior $G(33.2, 4.16)$ for the root age, with the mode at 774 Ma and 95% CI to be (550, 1090). Note that the gamma distribution $G(\alpha, \beta)$ has mean $\alpha/\beta$ and variance $\alpha/\beta^2$.

**Table 1**
Time constraints (Ma) on nodes used during the dating analysis (Fig. 1).

| Node number | Calibration | Fossil appearance |
|---|---|---|
| 1 | $G(33.2, 4.16)$ | Basal Monothalamids |
| 9 | $L(5)$ | *Ammodiscus* genus |
| 10 | $L(3.5)$ | *Cornuspira* genus |
| 13 | $L(1)$ | Peneroplidae family |
| 14 | $B(0.2, 0.5)$ | 0.2: Soritinae subfamily<br>0.5: Archaiasinae subfamily (sister group of Soritinae) |
| 20 | $L(4)$ | *Reophax* genus |
| 21 | $L(3.5)$ | *Trochammina* genus |
| 23 | $U(2)$ | Rotaliida order |
| 24 | $L(0.94)$ | Bolivinidae family |
| 30 | $L(0.90)$ | Rotaliidae family |
| 31 | $L(0.35)$ | Nummulitidae family (recent species) |
| 32 | $L(0.90)$ | *Stainforthia* genus |

*Note*: Node numbers refer to those in Fig. 1. The time unit is 100 Myrs. Four kinds of calibrations are used on the ages of nodes: $G(a, b)$ is the gamma distribution with mean $a/b$ and variance $a/b^2$; $B(a, b)$ is a pair of joint bounds $a < t < b$, implemented using equation (17) in Yang and Rannala (2006); $L(a)$ is the minimum-age (lower) bound, implemented using equation (26) in Inoue et al. (2010) with $p = 0.1$ and $c = 0.2$; and $U(b)$ is the maximum-age (upper) bound $t < b$, implemented using equation (16) in Yang and Rannala (2006).

The univariate calibration densities are multiplied and then truncated so that ancestral nodes are older than descendent nodes, leading to a joint distribution of ages of nodes with calibration information. This is multiplied by the prior density for ages of nodes without calibration information, specified using the birth-death process with species sampling, with the birth and death rates $\lambda = \mu = 2$ and the sampling fraction $\rho = 0.1$, representing a nearly flat kernel density (Yang and Rannala, 2006). As the truncation mentioned above can cause the effective prior used by the program to differ considerably from the apparent prior specified by the user (Inoue et al., 2010), we ran the program without the sequence data to generate the effective prior to make sure that it is reasonable judged by the fossil evidence.

The likelihood was calculated under the HKY85 + $\Gamma_5$ model, using both Felsenstein's (1981) exact algorithm and using the normal approximation (Thorne et al., 1998; Yang, 2006: figure 7.10). The two methods produced similar divergence time estimates. Our main results presented below were obtained using the approximate method. For the exact likelihood calculation, a gamma prior $G(6, 2)$ was assigned to the transition/transversion rate ratio $\kappa$ and a gamma prior $G(1, 1)$ was used for the gamma shape parameter $\alpha$ of the HKY85 + $\Gamma_5$ model.

The time unit is set at 100 Myr. To specify a gamma prior for the overall rate $\mu$, we fix its shape parameter at $\alpha = 1$ (which represents a diffuse prior), and obtain the scale parameter $\beta$ by fitting a molecular clock to the sequence data using point calibrations to estimate the prior mean ($=\alpha/\beta$). This led to the prior $\mu \sim G(1, 30)$, with the mean rate to be 0.033 per time unit or $3.3 \times 10^{-10}$ substitutions per site per year. Both the independent-rates and the correlated-rates models were used to accommodate variable rates between branches (Rannala and Yang, 2007). The two models produced similar time estimates (see below) so our main results are presented under the independent-rates model. Under the independent-rates model, a gamma prior is assigned on the variance of the logarithm of the rate: $\sigma^2 \sim G(1, 8)$: here again $\alpha = 1$ is chosen to represent a diffuse prior while the mean (1/8) is chosen as the reciprocal of prior mean of the root age, following the recommendation of the Multidivtime program (Thorne et al., 1998; Thorne and Kishino, 2002; Rutschmann, 2005). Under the correlated-rates model, the rate of the current branch depends on the rate of the ancestral branch, with the variance of the log rate to be $t\sigma^2$, if the current branch is $t$ time units later than the ancestral branch.

We refer to the above settings with the independent-rates model as the standard condition. We vary the prior or settings to evaluate the robustness of the analysis. In such a case, only one factor was altered at a time.

The MCMC was run for 300,000 iterations, with samples taken every five iterations, after a burnin of 10,000 iterations. At least two runs were launched to confirm the consistency between runs.

## 3. Results

### 3.1. Phylogenetic analysis

The ML and Bayesian trees are shown in Suppl. Figs. 1 and 2 respectively. The trees are unrooted, but the root is placed within monothalamiids to respect their traditional basal position (Pawlowski et al., 1999; Pawlowski and Holzmann, 2002; Ertan et al., 2004; Longet and Pawlowski, 2007) and the independent origins of two polythalamous clades (Pawlowski and Holzmann, 2002). The first polythalamous clade is composed of two orders: Rotaliida and Textulariida (see Fig. 1 and Suppl. Fig. 1). The monophyly of Rotaliida is found in both ML (Suppl. Fig. 1A and B) and Bayesian trees (Suppl. Fig. 1C and D), and with both the concatenated (Suppl. Fig. 1A and C) and partitioned data sets (Suppl. Fig. 1B and D), although the support in the PhyML (concatenated data set), RAxML (partitioned data set) and PhyloBayes (concatenated data set) trees is low, possibly due to particularly high evolutionary rates of *Ammonia* and *E. williamsoni* and low information content in the data. However, the posterior probability (PP) is high (1.00) in the MrBayes tree (partitioned data set). The monophyly of Textulariida is found in all trees and is consistent with the traditional morphology-based systematics (Loeblich and Tappan, 1987) and is thus retained in the input topology for MCMCTREE. The second polythalamous clade is composed of calcareous order Miliolida as well as some agglutinated genera (*Ammodiscus*, *Miliammina*). In all trees, *Miliammina* branches between *Ammodiscus* and *Bathysiphon*. This position is in disagreement with recent studies suggesting that this genus should be included in order Miliolida based on molecular, immunochemical and morphological features (Fahrni et al., 1997; Habura et al., 2006). The unconventional relationship may be the result of long-branch attraction since both *Ammodiscus* and Miliolida have long branches.

Therefore, this part of the tree was modified in the input tree for MCMCTREE, in order to respect palaeontological data concerning the timing of appearance in the fossil record of *Ammodiscus* (500 Ma) (node 9 in Fig. 1), *Cornuspira* (350 Ma) (node 10) and *Miliammina* (250 Ma) (node 11). It is worth mentioning that whatever is the branching order at the base of Miliolida, our phylogenetic analyses show consequently *Bathysiphon* as the sister group to Miliolida + *Ammodiscus* clade (node 8). Although the support for this relationship is low in the PhyML tree (66%), the RAxML, PhyloBayes and MrBayes trees strongly support the phylogenetic position (87%, 0.97 PP and 0.99 PP respectively). Thus, we hypothesize that the coiled tubular ancestor of *Ammodiscus* and miliolids originated from rectilinear tubular ancestor of recent *Bathysiphon*.

The relationships within the orders Rotaliida and Textulariida were not resolved in our ML and Bayesian analyses. To construct the input tree for MCMCTREE, we used the branching order within Rotaliida published by Schweizer et al. (2008), which distinguished three well supported clades. Palaeontological knowledge about the timing of appearance of *Reophax* (node 20) and *Trochammina* (node 21) was used to resolve the topology within Textulariida.

The rooted tree topology used for divergence time estimation by MCMCTREE is shown in Fig. 1. A few different input tree topologies were also used to evaluate the robustness of the posterior time estimates to the tree topology. We also examined the impact of the placement of the root.

### 3.2. Estimation of divergence times

#### 3.2.1. Likelihood ratio test of the molecular clock

We conducted the likelihood ratio test of the molecular clock hypothesis (Felsenstein, 1981) on the two partitions separately. The likelihood values were conducted with the BASEML program under the clock model and the no-clock models, without fossil calibrations (Yang, 2007). The clock model estimates 33 node ages on the rooted tree and the no-clock model estimates 65 branch lengths on the unrooted tree. Twice the log likelihood difference is compared with a $\chi^2$ distribution with 32° of freedom. For both partitions, the test rejected the molecular clock, with $p < 0.01$ for the coding-gene partition and $p < 0.001$ for the rDNA partition. The violation of the clock is also obvious from the ML and Bayesian branch lengths estimated under the no-clock model (Suppl. Fig. 1a and b).

#### 3.2.2. Estimates of divergence times under the standard condition

Fig. 1 shows the divergence times (chronogram) obtained under the independent-rates model. The posterior time estimates were in most cases in agreement with the fossil record. The radiation of Foraminifera was estimated to have occurred during the Cryogenian period; the posterior mean of the root age was 770 Ma with the 95% credibility interval (CI) to be (650–920). Under the correlated-rates model, the estimates were 700 Ma (610–815). These are younger than previous estimates (690–1150) obtained using the SSU only under a local-clock model (Pawlowski et al., 2003). Three main reasons could explain this difference and are presented in the following. (i) At this time, only partial SSU sequences were available. (ii) Fossil calibrations were only used for major foraminiferal radiations, which reduces the quantity of fossil information incorporated in the analysis. (iii) The method employed did not allow the rate to vary freely between lineages. Indeed, the substitution rate was estimated for multilocular species only and was then applied to all branches of the tree to calculate divergence times.

Most nodes relating monothalamid species (1, 2, 5 and 6), as well as node 7 between the monothalamous *Toxisarcon alba-Shinkaiya lindsayi* and polythalamous Textulariida + Rotaliida clade, were placed before the Early Cambrian. Those estimates support the classical idea that monothalamiids represented the vast majority of Foraminifera in the beginning of their radiation.

If the extremely fast-evolving *Ammodiscus* (see the ML and Bayesian trees in Suppl. Fig. 1) was removed, the posterior estimates of the root age became 736 Ma (600–890) under the independent-rates model, similar to estimates above. Thus removal of this species did not affect the time estimates by much.

At three major nodes, there appear to exist conflicts between the fossil calibrations and the molecular time estimates. The first concerns the origin of Rotaliida (node 23). The posterior age estimate was 190 (160–215) under the independent-rates model, with the upper limit of the CI to be even greater than the maximum fossil bound of 200 Ma. The conflict was even greater under the correlated-rates model, with the Rotaliida root dated to 240 (200–290) (Fig. 2A). When the fossil calibration was removed, almost all nodes within the Rotaliida were estimated to be much older than 200 Ma: the age of the Rotaliida root was estimated to be 380 Ma (300–460).

The second conflict concerns the origin of Miliolida (node 10). The posterior estimate, at 490 Ma (410–580), was much older than the fossil minimum bound of 350 Ma, implying a huge gap between the molecular date and the earliest fossil miliolids found. The estimate under the correlated-rates model was similar. Finally, the third conflict concerns calibration nodes in Textulariida. The posterior estimate for the origin (node 20 in Fig. 1) was 430 Ma (390–490), with the lower limit of the CI to be slightly younger than the minimum fossil bound of 400 Ma. When this fossil calibration was removed, the estimate became 396 Ma (352–470). Similarly the posterior estimates for the age of node 21 within textulariids, at which a minimum bound of 350 Ma was placed, were 380 (345–430), with the lower limit of the CI slightly lower than 350 Ma. When this fossil calibration was removed, the

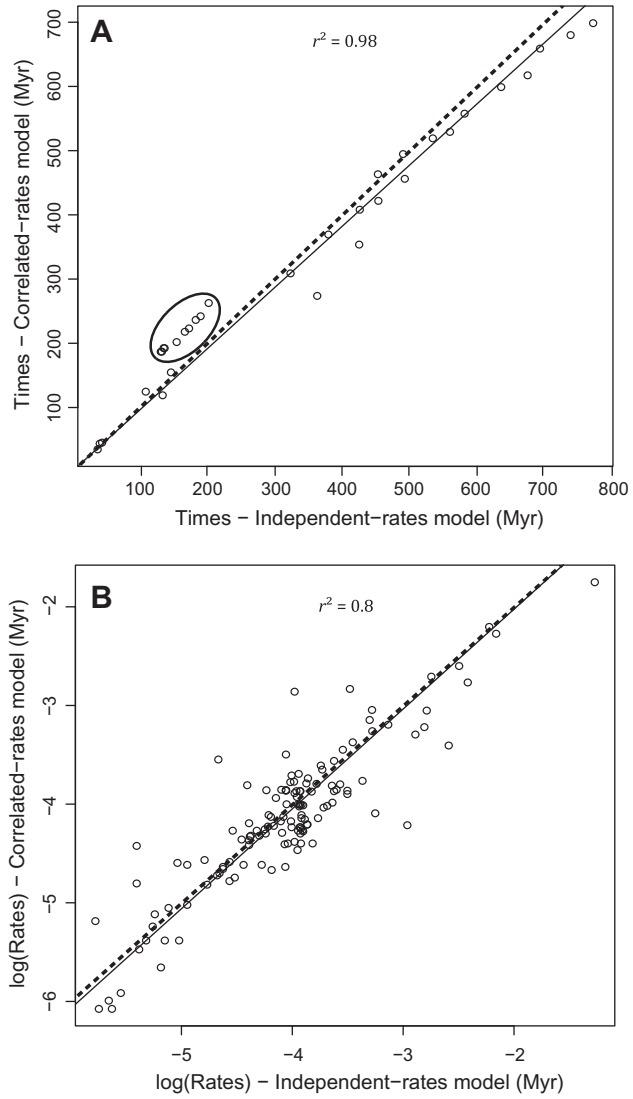



**Fig. 2.** Posterior means of times (A) and rates (B) estimated under the correlated-rates model plotted against those estimated under the independent-rates model. The solid line represents the regression line. The dashed line represents the $y = x$ line. The circle represents the violation of the 200 Myr maximum bound by most of the rotaliid nodes under the correlated-rates model.

estimates became 165 Ma (60–340), much younger than the fossil bound.

### 3.3. The impact of prior assumptions

We varied the standard condition to examine the impact of various factors on the posterior time estimation, such as the substitution model, data partitioning, the input tree topology, and the priors on times, rates and other parameters in the model.

#### 3.3.1. The influence of the rate-drift model

In Fig. 2 the posterior estimates of times and rates were compared between the independent-rates and correlated-rates models. Except for Rotaliida and deep nodes, the posterior time estimates were very similar between the two models. The age of the root was estimated to be 50 Myr younger with the correlated-rates model than under the independent-rates model. The rates were more different between the two analyses, but the correlation in the logarithm of the rates was quite high (with $r^2 = 0.79$).

#### 3.3.2. The influence of the prior on times

We varied the parameters (birth rate $\lambda$, death rate $\mu$, and sampling fraction $\rho$) in the birth-death process with species sampling used to specify the prior on the ages of the non-calibration nodes (Yang and Rannala, 2006). Besides the values used in the standard condition ($\lambda = \mu = 2$, $\rho = 0.1$) with a nearly flat kernel density, we also considered an L-shaped kernel ($\lambda = 1$, $\mu = 4$, $\rho = 0.1$), which produces trees with long internal branches, and an inverse L-shaped kernel ($\lambda = 4$, $\mu = 1$, $\rho = 0.0001$), which produces star-like trees with short internal branches. No difference was observed among those priors for most node ages, excepted for deep nodes (Suppl. Fig. 2A and B). The results were similar to those of Yang and Rannala (2006), who also found that parameters in the birth-death process with species sampling had little influence on the posterior time estimates.

We considered the impact of the prior on the ages of the calibration nodes. First, the influence of parameters $p$ and $c$ in the truncated Cauchy distribution for minimum bounds was examined. Inoue et al. (2010) analyzed three empirical datasets and found that both parameters, in particular $c$, had a strong influence on the posterior, with larger $p$ and $c$ pushing up estimates of all node ages. We used $c = 1$ in comparison with $c = 0.2$ in the standard condition (Suppl. Fig. 3A). All node ages became older in the prior. However, this effect did not persist in the posterior (Suppl. Fig. 3B). Second, we used different priors on the age of the root. In comparison with the gamma distribution $G(33.2, 4.16)$ in the standard condition, we also used joint minimum and maximum bounds $B(550, 1090)$. The results are presented in Suppl. Fig. 4. This prior produced very similar estimates for the root age to the gamma prior.

#### 3.3.3. The influence of the prior on rates

The effect of the gamma prior on the overall rate $\mu$ was examined. In the standard condition, $\mu \sim G(1, 30)$. We used $\mu \sim G(2, 60)$, so that the shape of the density is modified without changing the mean. This change in the prior produced very similar posterior estimates of times and rates (Suppl. Fig. 5).

In the standard condition, the rate-drift parameter $\sigma^2 \sim G(1, 8)$. We used $G(0.5, 8)$ and $G(10, 8)$ as well. Note that multiplying the shape parameter $\alpha$ by 0.5 reduces both the mean and variance by a half so that the rates are more homogeneous among lineages, and multiply $\alpha$ by 10 increases the mean and variance. The estimates (with those for $\alpha = 0.5$ shown in Suppl. Fig. 6) were very similar to those under the standard condition. Thus the prior assumption about the violation of the clock had little impact on the posterior time estimates. This result is in contrast with a previous study, which showed some impact of the prior on $\sigma^2$ (Inoue et al., 2010).

#### 3.3.4. The influence of the substitution model

We used the simple JC69 substitution model for comparison with HKY85 + $\Gamma_5$ assumed in the standard condition. The time estimates were very similar between the two models (Suppl. Fig. 7). This result emphasizes that the underestimation of sequence distances produced by too simple models is balanced by the multiple fossil calibrations used in the analysis, as it was previously noticed (Yang and Rannala, 2006). However, JC69 inferred much lower rates than HKY85 + $\Gamma_5$ (Suppl. Fig. 7B).

#### 3.3.5. The influence of the time unit used in the calculation

The birth-death process prior on times and the log-normal distribution of rates are not invariant to the change of the time unit. In theory, the results may differ when one changes the time unit from 100 Myr to 10 or 1000 Myr. We used all those three time units and found that the posterior time estimates were indistinguishable (Suppl. Fig. 8). We also reached the same conclusion using two

other datasets: the primate dataset used by Yang and Rannala (2006) and the fish Fish dataset of Inoue et al. (2010).

*3.3.6. Alternative models of time and rate priors*

We implemented two variations to the models of Yang and Rannala (2006) and Rannala and Yang (2007). First we replaced the birth-death kernel of Yang and Rannala (2006: equation 4) with a beta distribution when specifying the prior on times; in other words, the ages of the other nodes given the age of the root are order statistics of variables drawn from the beta distribution. Second, we implemented the gamma distribution of substitution rates in place of the log-normal under the independent-rates model. We found that the posterior estimates of times and rates were very similar to those obtained under the standard condition using the birth-death kernel for the prior on times and the log-normal distribution for the prior on rates (results not shown). The results suggest that the distributional details do not matter to the posterior time estimates in such modeling.

*3.3.7. The influence of the root placement and the tree topology*

Even if the root is likely to be located among monothalamids, its exact position is unknown because of the lack of closely related sister group to Foraminifera to be used as outgroups. In the topology used here, the position was somewhat arbitrarily chosen within basal monothalamids. Here we investigated the effect of the position of the root on the posterior estimates, by placing the root between all basal monothalamid species and the rest of the foraminifers. This change had little influence for almost all comparable nodes in the trees (Suppl. Fig. 9). Three nodes were estimated to be younger with the new root, but they are among the basal monothalamids, in the part of the tree where no fossil calibrations are available and where the times were hard to estimate. Their CIs were very wide and they overlap considerably between the analyses. Furthermore, the root age was estimated to be 700 Ma (600–842), compared with 770 Ma (650–920) with the previous root location. While the posterior mean was ~70 Myr younger, the posterior CIs overlap considerably.

Schweizer et al. (2008) recently proposed a well-supported phylogeny of rotaliids. They defined three major clades with strong statistical support. The present study focuses on the phylogeny of Foraminifera at a larger scale, so that many fast-evolving sites and insertions had to be deleted in the SSU alignment, preventing us to obtain an alignment supporting the Rotaliid topology found by Schweizer et al. (2008). To examine whether the Rotaliida topology is responsible for the violation of the 200 Ma maximum bound, we used a tree with the Rotaliid topology found by ML from the concatenated data (Suppl. Fig. 1A). The results were shown in Suppl. Fig. 10. All fossil calibrations within Rotaliida were respected by the posterior time estimates, and the Rotaliid root was dated to 200 Ma (180–225), indicating that the maximum bound is still violated. Thus the violation of the 200 Ma bound does not appear to be due to topological differences within Rotaliid.

*3.3.8. The influence of data partitioning*

We also analyzed the data as four partitions (the SSU rDNA vs. three coding genes), in comparison with the two-partition analysis under the standard condition. For the three protein coding genes, only the first and second codon positions were used, with 688, 658 and 802 sites respectively, and containing only 20, 16 and 8 species, respectively (see Suppl. Table 1). As with the two-partitions data set, the approximate and the exact likelihood calculations produced nearly identical results with the four-partitions data set. On the whole, posterior mean times are younger for the deepest nodes in the four-partitions analysis (Suppl. Fig. 11A). However, the CIs overlap widely between the two analyses. Finally, we also estimated divergence times using two mixed partitions (rDNA vs. amino acids). Suppl. Fig. 11B shows that posterior times are nearly identical to those obtained under the standard condition.

*3.4. Inter-species variations of evolutionary rates*

Suppl. Fig. 12 shows the two rategrams for the two partitions obtained under the standard condition. The rates for the coding-genes partition were quite homogeneous among all species and no important inter-species shifts were noticed (Suppl. Fig. 13A). However, 47% of sites in this partition are missing data. The rates were much more variable among lineages in the SSU partition, with monothalamiid and textulariid species having low rates and other parts of the tree, especially Miliolida, having high rates (Suppl. Fig. 12B). After the divergence of *Miliammina fusca* there was a fast rate acceleration, according to both the independent and correlated-rates models. This increase in evolutionary rates is accompanied by an increase in A + T content in miliolids (data not shown).

*3.5. Conflicts between coding-genes and SSU within the Rotaliida and Miliolida*

We analyzed the coding genes and the SSU data separately, to detect potential conflicts between the two sets of data and to examine which set is responsible for the violation of the 200 Ma maximum bound. Species having SSU sequences and no coding-gene sequence were removed from the SSU data set, so that only comparable nodes remained in the input topology. The posterior means of times estimated from the two datasets are shown in Fig. 3, which shows that four major discrepancies exist between coding genes and SSU. First, coding genes tend to predict younger times for the deeper nodes than the SSU. The root age estimate was 695 Ma (560, 860) from the genes and 750 Ma (580, 960) from the SSU. Interestingly, when the 200 Ma maximum bound for the origin of Rotaliids (node 23 in Fig. 1) was removed, the genes and SSU were in agreement concerning the times of deep nodes (Fig. 3B). Second, the two datasets gave very different age estimates for node 3 in Fig. 1, representing the divergence between *Crithionina delacai* and *Allogromia* + *Edaphoallogromia australica*: 260 Ma (120–520) for the genes and 610 (390–840) for the SSU. This node is far away from fossil calibrations and thus hard to date reliably. Third, the estimated ages of nodes for the most recent miliolids were older from the genes than from the SSU. For example, node 12 in Fig. 1, which groups *Quinqueloculina sp.* + *P. peruviana* with the taxon *P. pertusus* + *Sorites* + *Marginopora* + *A. hemprichii*, was estimated to be 400 Ma (170–490) from the genes and 160 Ma (100–300) from the SSU. The extremely long branch between nodes 11 and 12 (Fig. 1) for the SSU (Suppl. Fig. 12B) appears to explain why MCMCTREE infers a wide time interval between the two nodes and a young age for the most recent miliolids (node 12). Fourth, as with the most recent miliolids, coding-genes tend to predict older nodes within the Rotaliida. The origin of Rotaliida is dated at 190 Ma (160, 215) with the genes and at 160 Ma (120, 200) with SSU. Thus, SSU seemed to be in agreement with the 200 Ma maximum bound while the genes were not. When the maximum bound was removed, the estimated age for the origin of Rotaliida remained reasonable with SSU, at 230 Ma (120, 400), but questionable with the genes, at 380 Ma (300, 450) (Fig. 3B).

*3.6. Infinite-sites plot*

For a fixed set of fossil calibrations, the errors in the posterior time estimates will not approach zero when the amount of sequence data increases. Instead the joint posterior distribution will become one-dimensional, and as a result, the posterior CI widths will be linear with the posterior means (Yang and Rannala, 2006). One can then plot the posterior CI width against the

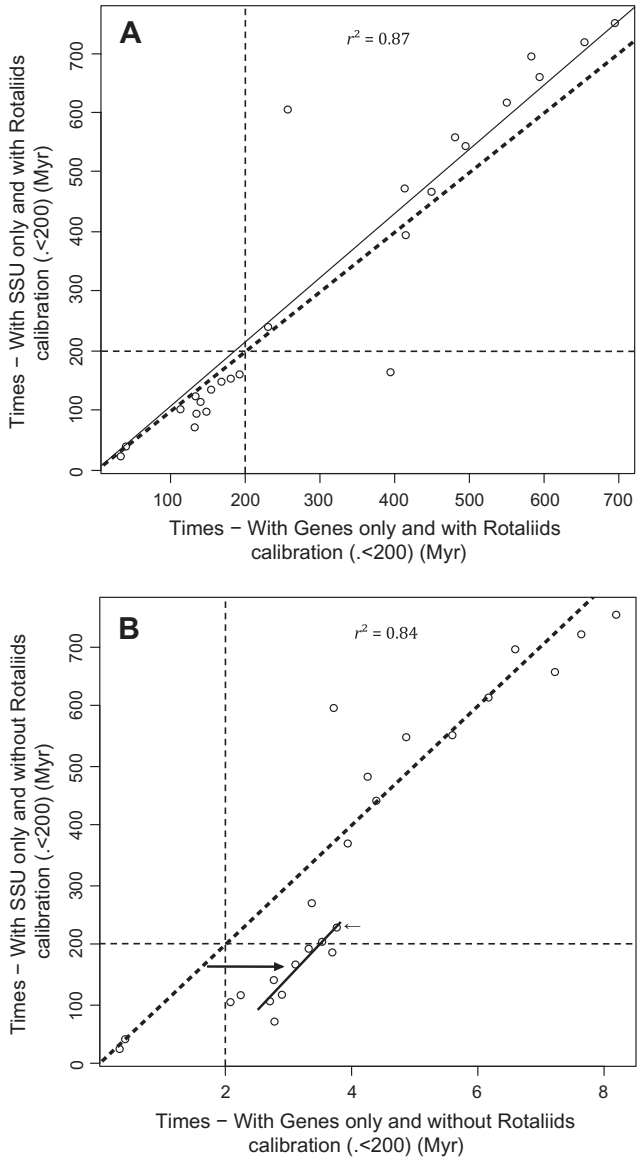



**Fig. 3.** Posterior means of times estimated using the protein-coding genes plotted against the posterior means of times estimated using the SSU, either with maximum bound on the root of Rotaliids used (A) or removed (B). The thin dashed lines represents the 200 Myr maximum bound on the root of rotaliids. The thick arrow represents the shift of time estimates of the rotaliid nodes.

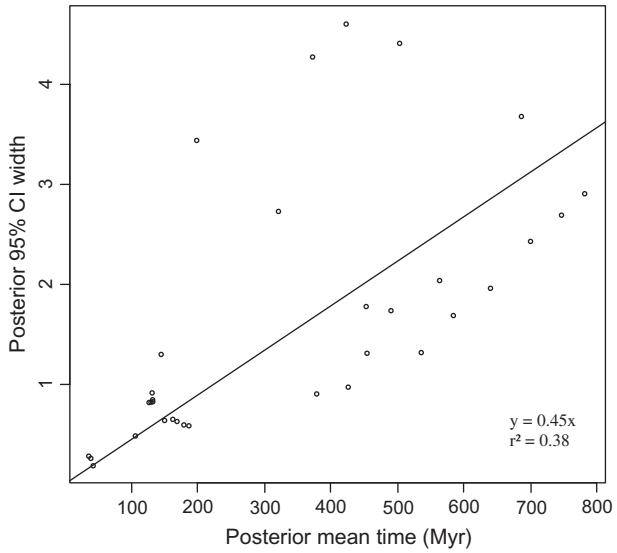



**Fig. 4.** The infinite-sites plot in which the widths of the posterior 95% CIs are plotted against the posterior means of divergence times. Six nodes (all above the regression line) with very wide CIs are highlighted by asterisks in Fig. 1B. Three of them belong to Monothalamiida, where no fossil calibration is available.

posterior mean of node ages to assess whether the amount of sequence data is near saturation or additional sequence data are likely to increase the precision of estimation. The infinite-sites plot for the two-partition data is shown in Fig. 4. The slope (0.45) means that every 1 Myr of divergence adds about 0.45 Myr of uncertainty in the posterior CI. The considerable scatter in the plot (with $r^2 = 0.38$) suggests that the sequence data are rather limited and sequencing new genes or adding species will very likely lead to more precise estimates. It is noteworthy that three of the most-poorly dated nodes are ancestors of monothalamiid species, in the part of the tree with no fossil calibrations.

## 4. Discussion

While molecular data has not yet completely resolved the phylogenetic relationships or produced definitive estimates of divergence times in Eukaryotes or Foraminifera, our time estimates are much more precise than in previous studies (Berney and Pawlowski, 2006; Douzery et al., 2004). This is clearly due to our use of multiple fossil calibrations and the expanded molecular dataset. Our estimates suggest that the radiation of Foraminifera occurred around 750 Ma, between approximately 650 and 900 Ma. This estimate is much younger and more precise than the estimate of Pawlowski et al. (2003), who proposed, on the basis of an SSU-only analysis, a CI of (690–1150). Our new estimate is more in line with the non-basal phylogenetic position of Foraminifera among Rhizaria in SSU phylogenies (Pawlowski and Burki, 2009). Our estimates support the hypothesis that all eukaryotic super-groups emerged during the Neo-Proterozoic (Berney and Pawlowski, 2006; Douzery et al., 2004). As the oldest fossils unequivocally attributed to Foraminifera are dated from the Early Cambrian, the results imply that the evolutionary history of this group includes a long non-fossilized period. According to our estimates, almost all divergence events among monothalamids occurred before the Cambrian explosion (Fig. 1) during the Cryogenian period (Neo-Proterozoic). Those results indicate that the Cryogenian oceanic fauna was composed of monothalamous foraminiferal species and that Foraminifera are important to the Neo-Proterozoic protistan communities and the difficulties in finding fossil traces of foraminifers during this period could be due to non-preservation of these specimens in rocks. Some microfossils, dated from the Neo-Proterozoic, were previously discovered (Porter and Knoll, 2000; Rasmussen et al., 2002). Nevertheless, it was complicated to attribute them to particular lineages because of too simple morphologies, which prompted Pawlowski et al. (2003) to suggest a reevaluation of the interpretations in these studies concerning whether these fossils represent unilocular foraminifers. The results presented here strongly support such a conclusion.

On the whole, the divergence times estimated using the molecular data are in agreement with the fossil record. The soft bounds implemented in the MCMCTREE program appear useful for detecting possible conflicts between the fossils and the molecules, as indicated by the posterior CIs going beyond the specified bounds. We noted that the 400 Ma minimum bound for the origin of Textulariids was slightly violated while the 350 Ma minimum bound within this group was likely to be inappropriate (see Section 3). This could be explained by the fact that only *Reophax* possesses gene sequences in the coding-genes partition and that the SSU

evolutionary rates for the four textulariid species are really low (Suppl. Fig. 12B). By allowing the algorithm to propose dates that do not conform with fossil calibrations during the MCMC calculation, MCMCTREE appears to have detected a more general problem concerning the uncertain status of the genus *Trochammina*. Theoretically, this is a very old genus, which appeared in Carboniferous (350 Ma) (Haynes, 1981) but its taxonomic definition is unclear and it is possible that the genus is not monophyletic. Thus we cannot rule out the possibility that the *Trochammina sp.* that was sequenced is not so old and that its divergence with *Eggerelloides scabrum* + *Textularia sagittula* occurred much later. In agreement with this, both *Textularia* and *Eggerelloides* are relatively recent genera that appeared in Paleocene and Holocene, respectively.

Furthermore, the maximum bound of 200 Ma for the radiation of Rotaliida is violated by the molecular time estimates. Under the correlated-rates model, both the mean and the CI upper limit are older than 200 Ma (see Section 3). Two reasons may explain the results. First, the fossil calibration may not represent the true evolutionary history of this group, which may have diverged from aragonitic lineage earlier in the Triassic (Haynes, 1981). Second, the dating method may not handle fast rates within Rotaliida and as a consequence overestimates the time of divergence. The SSU data set was particularly subjected to strong inter-species variation of rates within Rotaliida (Suppl. Fig. 12B), with particularly high rates for *Ammonia* and *E. williamsoni*. As it appeared that it was the coding-genes that are not in agreement with the fossil calibration, further investigations are needed to understand the impact of fast rates of coding-genes within Rotaliida on posterior time estimation. We can conclude that, although estimating the divergence time of this group is non-trivial, an origin of modern Rotaliida during the Jurassic is still plausible in respect to the SSU data set. The third conflict between molecular and fossil data concerns the origin of Miliolida. In the fossil record, miliolids appear during the Carboniferous around 350 Ma, with the first species assigned to the family Cornuspiridae (Haynes, 1981). The genus *Miliammina* emerged in the fossil record during the Triassic (from 251 to 200 Ma). Our main analysis places the radiation of Miliolida around 490 Ma, under both the independent- and correlated-rates models and with analogous results between SSU and coding-genes when these are analyzed separately. These results would imply that a period of 100–140 Myr of the miliolids history did not leave fossil traces in sedimentary rocks. Moreover, our estimates place the time of divergence between *M. fusca* and the rest of miliolids during the Devonian or Silurian between 390 and 450 Ma, depending on the data set analyzed. Thus a similar conclusion can be drawn for *Miliammina*, with a non-fossilized period of 200 Myr.

The following factors may explain why our results are not compatible with the classical scheme followed by palaeontologists. (i) Both the branches leading to *Cornuspira* and node 12 are characterized by rapid evolutionary rates in SSU (Suppl. Fig. 12B). MCMCTREE may encounter difficulties to handle those extreme high rates. (ii) The origin of Miliolida corresponds to the apparition of calcareous tests. Consequently, the early *Cornuspira* could be very weakly calcified and therefore not preserved in the fossil record. (iii) The common ancestor of Miliolida was actually still agglutinated and *Cornuspira* developed a calcified test independently from the other miliolids. However, this third hypothesis does not appear to be plausible for the reason that *Cornuspira* has a typical miliolid porcellaneous wall, and that it is not parsimonious to imagine that the formation of such wall originated more than once. Similarly, the very early divergence of *Miliammina* is unexpected since this genus has particular type of chambers winding in different planes, typical for *Quinqueloculina* and other Miliolacea that appeared in Jurassic. It is unlikely that foraminifers with such characteristic type of test have not been noticed in the fossil record by palaeontologists.

Accurately estimating divergence times during the Neo-Proterozoic among Rhizaria is necessary to understand the speciation dynamic of early Eukaryotes. In particular, inferring the time of the radiation of Radiolaria, sister-group of Foraminifera according to recent results (Burki et al., 2010), will be of great interest to understand the origin of Foraminifera. Thus, accumulation of more sequence data will most likely lead to more precise time estimates, particularly in groups of unilocular species where no fossil calibration is available, as indicated by the infinite-sites plot. Genetic data less prone to evolutionary rate variation than SSU will be especially valuable in resolving the conflicts between fossil and molecular dates observed in this study.

## Acknowledgments

We are grateful to an anonymous referee for many constructive comments. The authors would like to sincerely thank Yurika Ujiie for providing some of the gene sequences used in this study. M.G. was supported by the École Normale Supérieure (ENS) of Lyon. J.P. was supported by the Swiss National Science Foundation Grant 31003A-125372. Z.Y. gratefully acknowledges the support of K.C. Wong Education Foundation, Hong Kong.

## Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2011.06.008.

## References

Bell, C.D., Donoghue, M.J., 2005. Dating the dipsacales: comparing models, genes, and evolutionary implications. Am. J. Bot. 92, 284–296.
Berney, C., Pawlowski, J., 2006. A molecular time-scale for eukaryote evolution recalibrated with the continuous microfossil record. Proc. R. Soc. B 273, 1867–1872.
Burki, F., Kudryavtsev, A., Matz, M.V., Aglyamova, G.V., Bulman, S., Fiers, M., Keeling, P.J., Pawlowski, J., 2010. Evolution of Rhizaria: new insights from phylogenomic analysis of uncultivated protists. BMC Evol. Biol. 10, 377.
Culver, S.J., 1991. Early cambrian Foraminifera from West Africa. Science 254, 689–691.
Douzery, E.J.P., Snell, E.A., Bapteste, E., Delsuc, F., Philippe, H., 2004. The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? Proc. Natl. Acad. Sci. USA 101, 15386–15391.
Edgar, R.C., 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinform. 5, 113.
Ertan, K.T., Hemleben, V., Hemleben, C., 2004. Molecular evolution of some selected benthic foraminifera as inferred from sequences of the small subunit ribosomal DNA. Mar. Micropaleontol. 53, 367–388.
Fahrni, J., Pawlowski, J., Richardson, S., Debenay, J.P., Zaninetti, L., 1997. Actin suggests Miliammina fusca (Brady) is related to porcellaneous rather than to agglutinated foraminifera. Micropaleontology 43, 211–214.
Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17, 368–376.
Gaucher, C., Sprechmann, P., 1999. Upper Vendian skeletal fauna of the Arroyo del Soldado Group, Uruguay. Beringeria 23, 55–91.
Gouy, M., Guindon, S., Gascuel, O., 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. 27, 221–224.
Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696–704.
Habura, A., Goldstein, S.T., Parfrey, L.W., Bowser, S.S., 2006. Phylogeny and ultrastructure of *Miliammina fusca*: evidence for secondary loss of calcification in a miliolid foraminifer. J. Eukaryot. Microbiol. 53, 204–210.
Hasegawa, M., Thorne, J.L., Kishino, H., 2003. Time scale of eutherian evolution estimated without assuming a constant rate of molecular evolution. Genes Genet. Syst. 78, 267–283.
Haynes, J.R., 1981. Foraminifera. John Wiley & Sons, New York, NY, 433 p..
Holzmann, M., Hohenegger, J., Hallock, P., Piller, W.E., Pawlowski, J., 2001. Molecular phylogeny of large miliolid foraminifera (Soritacea Ehrenberg 1839). Mar. Micropaleontol. 43, 57–74.
Holzmann, M., Habura, A., Giles, H., Bowser, S.S., Pawlowski, J., 2003. Freshwater foraminiferans revealed by analysis of environmental DNA samples. J. Eukaryot. Microbiol. 50, 135–139.
Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogeny. Bioinformatics 17, 754–755.

Inoue, J., Donoghue, P.C.J., Yang, Z., 2010. The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. Syst. Biol. 59, 74–89.
Keeling, P.J., Burger, G., Durnford, D.G., Lang, B.F., Lee, R.W., Pearlman, R.E., Roger, A.J., Gray, M.W., 2005. The tree of eukaryotes. Trends Ecol. Evol. 20, 670–676.
Kostka, M., Uzlikova, M., Cepicka, I., Flegr, J., 2008. SlowFaster, a user-friendly program for slow-fast analysis and its application on phylogeny of Blastocystis. BMC Bioinform. 9, 341.
Lartillot, N., Lepage, T., Blanquart, S., 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25, 2286–2288.
Lejzerowicz, F., Pawlowski, J., Fraissinet-Tachet, L., Marmeisse, R., 2010. Molecular evidence for widespread occurrence of Foraminifera in soils. Environ. Microbiol. 12, 2518–2526.
Loeblich Jr., A.R., Tappan, H., 1987. Foraminiferal Genera and Their Classification. Van Nostrand Reinhold Company, New York, 2047 pp.
Longet, D., Pawlowski, J., 2007. Higher-level phylogeny of Foraminifera inferred from the RNA polymerase II (RPB1) gene. Eur. J. Protistol. 43, 171–177.
McIlroy, D., Green, O.R., Brasier, M.D., 2001. Paleobiology and evolution of the earliest agglutinated Foraminifera: platysolenites, spirosolenites and related forms. Lethaia 34, 13–29.
Pawlowski, J., Bolivar, I., Guiard-Maffia, J., Gouy, M., 1994. Phylogenetic position of Foraminifera inferred from LSU rRNA gene sequences. Mol. Biol. Evol. 11, 929–938.
Pawlowski, J., Bolivar, I., Fahrni, J.F., de Vargas, C., Gouy, M., Zaninetti, L., 1997. Extreme differences in rates of molecular evolution of Foraminifera revealed by comparison of ribosomal DNA sequences and the fossil record. Mol. Biol. Evol. 14, 498–505.
Pawlowski, J., Bolivar, I., Fahrni, J.F., de Vargas, C., Bowser, S.S., 1999. Molecular evidence that Reticulomyxa filosa is a freshwater naked foraminifer. J. Eukaryot. Microbiol. 46, 612–617.
Pawlowski, J., Holzmann, M., 2002. Molecular phylogeny of Foraminifera – a review. Eur. J. Protistol. 38, 1–10.
Pawlowski, J., Holzmann, M., Berney, C., Fahrni, J., Gooday, A.J., Cedhagen, T., Habura, A., Bowser, S.S., 2003. The evolution of early Foraminifera. Proc. Natl. Acad. Sci. USA 100, 11491–11498.
Pawlowski, J., Burki, F., 2009. Untangling the phylogeny of amoeboid protists. J. Eukaryot. Microbiol. 56, 16–25.
Porter, S.M., Knoll, A.H., 2000. Testate amoebae in the Neoproterozoic era: evidence from vase-shaped microfossils in the Chuar Group, Grand Canyon. Paleobiology 26, 360–385.
Rannala, B., Yang, Z., 2007. Inferring speciation times under an episodic molecular clock. Syst. Biol. 56, 453–466.
Rasmussen, B., Bengtson, S., Fletcher, I.R., McNaughton, N.J., 2002. Discoidal impressions and trace-like fossils more than 1200 Million years old. Science 296, 1112–1115.
Ronquist, F., Huelsenbeck, J.P., 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.
Rutschmann, F., 2005. Bayesian molecular dating using PAML/multidivtime. A step-by-step manual. University of Zurich, Switzerland. <http://www.plant.ch>.
Schweizer, M., Pawlowski, J., Kouwenhoven, T.J., Guiard, J., van der Zwaan, B., 2008. Molecular phylogeny of Rotaliida (Foraminifera) based on complete small subunit rDNA sequences. Mar. Micropaleontol. 66, 233–246.
Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690.
Thorne, J.L., Kishino, H., Painter, I.S., 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15, 1647–1657.
Thorne, J.L., Kishino, H., 2002. Divergence time and evolutionary rate estimation with multilocus data. Syst. Biol. 51, 689–702.
Ujiié, Y., Kimoto, K., Pawlowski, J., 2008. Molecular evidence for an independent origin of modern triserial planktonic foraminifera from benthic ancestors. Mar. Micropaleontol. 69, 334–340.
Yang, Z., 1996. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42, 587–596.
Yang, Z., 2006. Computational Molecular Evolution. Oxford University Press, Oxford, England.
Yang, Z., Rannala, B., 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol. Biol. Evol. 23, 212–226.
Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591.

# Bibliographie


Abby SS, Tannier E, Gouy M, and Daubin V. 2012. Lateral gene transfer as a support for the tree of life. *Proc Natl Acad Sci U S A* 109 :4962–4967.

Adachi J and Hasegawa M. 1996. MOLPHY version 2.3 : programs for molecular phylogenetics based on maximum likelihood. :1–150. *Comput Sci Monogr* 28 :1–150.

Anisimova M, Liberles DA, Philippe H, Provan J, Pupko T, and von Haeseler A. 2013. State-of the art methodologies dictate new standards for phylogenetic analysis. *BMC Evol Biol* 13 :161.

Bentley SD and Parkhill J. 2004. Comparative genomic structure of prokaryotes. *Annu Rev Genet* 38 :771–792.

Bershtein S, Goldin K, Tawfik DS, et al.. 2008. Intense neutral drifts yield robust and evolvable consensus proteins. *J Mol Biol* 379 :1029–1044.

Blanquart S and Lartillot N. 2006. A Bayesian Compound Stochastic Process for Modeling Nonstationary and Nonhomogeneous Sequence Evolution. *Mol Biol Evol* 23 :2058–2071.

Blanquart S and Lartillot N. 2008. A Site- and Time-Heterogeneous Model of Amino Acid Replacement. *Mol Biol Evol* 25(5) :842–858.

Bos KI, Schuenemann VJ, Golding GB, et al.. 2011. A draft genome of Yersinia pestis from victims of the Black Death. *Nature* 478 :506–510.

Boussau B, Blanquart S, Necsulea A, Lartillot N, and Gouy M. 2008. Parallel Adaptation to High Temperature in the Archaean Eon. *Nature* 456 :942–945.

Boussau B and Gouy M. 2006. Efficient likelihood computations with nonreversible models of evolution. *Syst Biol* 55 :756–768.

Brent R et al.. 1973. *Algorithms for minimization without derivatives*. Prentice-Hall, Englewood Clifts.

Bromham L, Rambaut A, Fortey R, Cooper A, and Penny D. 1998. Testing the cambrian explosion hypothesis by using a molecular dating technique. *Proc Natl Acad Sci U S A* 95 :12386–12389.

Cao Y, Adachi J, Janke A, et al.. 1994. Phylogenetics relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins : instability of a tree based on a single gene. *J Mol Evol* 39 :519–527.

Cole MF and Gaucher EA. 2011. Utilizing natural diversity to evolve protein function : applications towards thermostability. *Curr Opin Chem Biol* 15 :399–406.

Dayhoff MO, Schwartz RM, and Orcutt BC. 1978. A model of evolutionary change in proteins. In *Atlas of protein sequence and structure* volume 5 pages 345–352. National Biomedical Research Foundation.

Denault M, Pelletier JN, et al.. 2007. Protein library design and screening : working out the possibilities. *Methods Mol Biol* 352 :127–154.

Doolittle R and Blombäck B. 1964. Amino-Acid Sequence Investigations of Fibrinopeptides from Various Mammals : Evolutionary Implications. *Nature* 202 :147–152.

Drummond AJ, Ho SY, Phillips MJ, and Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *Plos Biol* 4 :e88.

Dutheil J and Boussau B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol Biol* 8 :255.

Dutheil J, Gaillard S, Bazin E, Glémin S, Ranwez V, Galtier N, and Belkhir K. 2006. Bio++ : a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinformatics* 7 :188.

Felsenstein J. 1981. Evolutionary Trees from DNA Sequences : A Maximum Likelihood Approach. *J Mol Evol* 17 :368–376.

Felsenstein J. 2004. *Inferring phylogenies*. Sunderland (ma) : Sinauer associates. edition.

Felsenstein J, Churchill GA, et al.. 1996. A hidden Markov model approach to variation among sites in rate of evolution. *Mol Biol Evol* 13 :93–104.

Fitch WM. 1971. Toward defining course of evolution–minimum change for a specific tree topology. *Syst Zool* 20 :406–416.

Foster PG. 2004. Modeling compositional heterogeneity. *Syst Biol* 53 :485–495.

Foster PG and Hickey DA. 1999. Compositional bias may affect both DNA-based and Protein-based phylogenetic reconstructions. *J Mol Evol* 48 :284–290.

Fox GE, Magrum LJ, Balch WE, Wolfe RS, and Woese CR. 1977. Classification of methanogenic bacteria by 16S ribosomal characterization. *Proc Natl Acad Sci U S A* 74 :4537–4541.

Galtier N. 2001. Maximum-Likelihood Phylogenetic Analysis Under a Covarion-like Model. *Mol Biol Evol* 18 :866–873.

Galtier N and Duret L. 2007. Adaptation or biased gene conversion ? Extending the null hypothesis of molecular evolution. *Trends Genet* 23 :273–277.

Galtier N and Gouy M. 1998. Inferring Pattern and Process : Maximum-Likelihood Implementation of a Nonhomogeneous Model of DNA Sequence Evolution for Phylogenetic Analysis. *Mol Biol Evol* 15 :871–879.

Galtier N, Tourasse N, and Gouy M. 1999. A Nonhyperthermophilic Common Ancestor to Extant Life Forms. *Science* 283 :220–221.

Gautier C. 2000. Compositional bias in DNA. *Curr Opin Genet Dev* 10 :656–661.

Gogarten JP, Kibak H, Dittrich P, Taiz L, Bowman EJ, Bowman BJ, Manolson MF, Poole RJ, Date T, Oshima T, Konishi J, Denda K, and Yoshida M. 1989. Evolution of the vacuolar H+-ATPase : Implications for the origin of eukaryotes. *Proc Natl Acad Sci U S A* 86 :6661–6665.

Goldman N and Yang Z. 1994. A codon-based model for nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* 11 :725–736.

Graur D and Martin W. 2004. Reading the entrails of chickens : molecular timescales of evolution and the illusion of precision. *Trends Genet* 20 :80–96.

Green RE, Krause J, Briggs AW, et al.. 2010. A draft sequence of the Neandertal genome. *Science* 328 :710–722.

Groussin M, Boussau B, and Gouy M. 2013a. A Branch-Heterogeneous Model of Protein Evolution for Efficient Inference of Ancestral Sequences. *Syst Biol* 62 :523–538.

Groussin M, Guéguen L, Boussau B, Gouy M, and Lartillot N. 2013b. Efficient modeling of protein site-heterogeneities with empirical mixtures of profiles. *Chapitre 2, section 2 de ce manuscrit de thèse*.

Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, Bigot T, Fournier D, Pouyet F, Cahais V, Bernard A, Scornavacca C, Nabholz B, Haudry A, Dachary L, Galtier N, Belkhir K, and Dutheil JY. 2013. Bio++ : Efficient Extensible Libraries and Tools for computational molecular evolution. *Mol Biol Evol* 30 :1745–1750.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, and Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies : assessing the performance of PhyML 3.0. *Syst Biol* 59 :307–321.

Hanson-Smith V, Kolaczkowski B, and Thornton JW. 2010. Robustness of Ancestral Sequence Reconstruction to Phylogenetic Uncertainty. *Mol Biol Evol* 27 :1988–1999.

Hastings WK. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57 :97–109.

Heath TA. 2012. A hierarchical bayesian model for calibrating estimated of species divergence times. *Syst Biol* 61 :793–809.

Hedges SB, Parker PH, Sibley CG, and Kumar S. 1996. Continental breakup and the ordinal diversification of birds and mammals. *Nature* 16 :226–229.

Henikoff S and Henikoff J. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 89 :10915–10919.

Hershberg R and Petrov DA. 2010. Evidence That Mutation Is Universally Biased towards AT in Bacteria. *Plos Genet* 6 :e1001115.

Hildebrand F, Meyer A, and Eyre-Walker A. 2010. Evidence of Selection upon Genomic GC-Content in Bacteria. *Plos Genet* 6.

Huelsenbeck JP, Bollback JP, and Levine AM. 2002. Inferring the Root of a Phylogenetic Tree. *Syst Biol* 51 :32–43.

Huelsenbeck JP, Ronquist F, Nielsen R, and Bollback JP. 2001. Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology. *Science* 294 :2310–2314.

Iwabe N, Kuma K, Hasegawa M, Osawa S, and Miyata T. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc Natl Acad Sci U S A* 86 :9355–9359.

Jermiin LS, Ho SYW, Ababneh F, Robinson J, and Larkum AWD. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Syst Biol* 53 :638–643.

Jones DT, Taylor WR, and Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8 :275–282.

Jukes TH and Cantor CR. 1969. Evolution of protein molecules. In *Mammalian protein metabolism* pages 21–123. Academic Press h. n. munro edition.

Kimura M. 1968. Evolutionary Rate at the Molecular Level. *Nature* 217 :624–626.

Kiraga J, Mackiewicz P, Mackiewicz D, Kowalczuk M, Biecek P, Polak N, Smolarczyk K, Dudek MR, and Cebrat S. 2007. The relationships between the isoelectric point and : length of proteins, taxonomy and ecology of organisms. *BMC Genomics* 8 :163.

Kumar S. 2005. Molecular clocks : four decades of evolution. *Nat Rev Genet* 6 :654–662.

Lartillot N and Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol* 21 :1095–2004.

Le SQ, Dang CC, and Gascuel O. 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol Biol Evol* 29 :2921–2936.

Le SQ and Gascuel O. 2008. An Improved General Amino Acid Replacement Matrix. *Mol Biol Evol* 25 :1307–1320.

Le SQ and Gascuel O. 2010. Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. *Syst Biol* 59 :277–287.

Le SQ, Gascuel O, and Lartillot N. 2008a. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24 :2317–2323.

Le SQ, Lartillot N, and Gascuel O. 2008b. Phylogenetic mixture models for proteins. *Phil. Trans. R. Soc. Lond. B* 363 :3965–3976.

Lehmann M, Loch C, Middendorf A, Studer D, Lassen SF, Pasamontes L, van Loon AP, Wyss M, et al.. 2002. The consensus concept for thermostability engineering of proteins : further proof of concept. *Protein Eng* 15 :403–411.

Lewis F, Butler A, and Gilbert L. 2011. A unified approach to model selection using the likelihood ratio test. *Methods in Ecology and Evolution* 2 :155–162.

Lobry JR and Necsulea A. 2006. Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes. *Gene* 385 :128–136.

Lopez P, Forterre P, and Philippe H. 1999. The Root of the Tree of Life in the Light of the Covarion Model. *J Mol Evol* 49 :496–508.

Madern D, Ebel C, and Zaccai G. 2000. Halophilic adaptation of enzymes. *Extremophiles* 4 :91–98.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, and Teller E. 1953. Equations of state calculations by fast computing machines. *J Chemical Physics* 21 :1087–1091.

Muse SV and Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol* 11 :715–724.

Naya H, Romero H, Zavala A, Alvarez B, and Musto H. 2002. Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. *J Mol Evol* 55 :260–264.

Ness JE, Kim S, Gottman A, Pak R, Krebber A, Borchert TV, Govindarajan S, Mundorff EC, Minshull J, et al.. 2002. Synthetic shuffling expands functional protein diversity by allowing amino acids to recombine independently. *Nat Biotechnol* 20 :1251–1255.

Nielsen R and Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148(3) :929–936.

Pagel M. 1999. Inferring the historical patterns of biological evolution. *Nature* 401 :877–884.

Pagel M and Meade A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst Biol* 53 :571–581.

Pagel M, Meade A, and Barker D. 2004. Bayesian Estimation of ancestral character states on phylogenies. *Syst Biol* 53 :673–684.

Paul S, Bag SK, Das S, Harvill ET, and Dutta C. 2008. Molecular signature of hypersaline adaptation : insights from genome and proteome composition of halophilic prokaryotes. *Genome Biol* 9 :R70.

Pauling L and Zuckerkandl E. 1963. Chemical Paleogenetics : Molecular "Restoration Studies" of Extinct Forms of Life. *Acta Chem Scand* 17 :S9–S16.

Perrière G and Brochier-Armanet C. 2010. *Concepts et méthodes en phylogénie moléculaire*. Paris springer edition.

Philippe H, Derelle R, Lopez P, Pick K, Borchiellini C, Bourry-Esnault N, Vacelet J, Renard E, Houliston E, Quéinnec E, Silva CD, Wincker P, Guyader HL, Leys S, Jackson DJ, Schreiber F, Erpenbeck D, Morgenstern B, Wörheide G, and Manuel M. 2009. Phylogenomics revives traditional views on deep animal relationships. *Curr Biol* 19 :706–712.

Philippe H and Forterre P. 1999. The Rooting of the Universal Tree of Life Is Not Reliable. *J Mol Evol* 49 :509–523.

Philippe H and Roure B. 2011. Difficult phylogenetic questions : more data, maybe ; better methods, certainly. *BMC Biology* 9 :91.

Pina-Aguilar RE, Lopez-saucedo J, Sheffield R, Ruiz-Galaz LI, Barroso-Padilla J, and Guitiérrez-Gutiérrez A. 2009. Revival of extinct species using nuclear transfer : hope for the Mammoth, true for the pyrenean ibex, but is it time for "Conservation Cloning" ? *Cloning and Stem Celles* 11 :341–346.

Pupko T, Doron-Faigenboim A, Liberles DA, and Cannarozzi GM. 2007. Probabilistic models and their impact on the accuracy of reconstructed ancestral protein sequences. In *Ancestral Sequence Reconstruction* pages 43–57. Oxford University Press.

Pupko T, Pe'er I, Hasegawa M, Graur D, and Friedman N. 2002. A Branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites : application to the evolution of five gene families. *Bioinformatics* 18 :1116–1123.

Pupko T, Pe'er I, Shamir R, and Graur D. 2000. A Fast Algorithm for Joint Reconstruction of Ancestral Amino Acid Sequences. *Mol Biol Evol* 17 :890–896.

Rannala B and Yang Z. 2007. Inferring speciation times under an episodic molecular clock. *Syst Biol* 56 :453–466.

Reich D, Green RE, Kircher M, et al.. 2010. Genetic history of an archaic hominin group from Denisova cave in Siberia. *Nature* 468 :1053–1060.

Rocha EP, Touchon M, and Feil EJ. 2006. Similar compositional biases are caused by very different mutational effects. *Genome Res* 16 :1537–1547.

Rodrigue N, Philippe H, and Lartillot N. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc Natl Acad Sci U S A* 107 :4629–4634.

Sankoff D. 1975. Minimal mutation trees of sequences. *SIAM J Appl Math* 28 :35–42.

Schluter. 1995. Uncertainty in ancient phylogenies. *Nature* 377 :108–109.

Schneider R, de Daruvar A, and Sander C. 1997. The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res* 25 :226–230.

Shapiro B, Rambaut A, and Drummond AJ. 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol Biol Evol* 23 :7–9.

Singer GAC and Hickey DA. 2003. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene* 317 :39–47.

Sueoka N. 1962. On the genetic basis od variation and heterogeneity of DNA base composition. *Proc Natl Acad Sci U S A* 15 :582–592.

Sullivan J, Swofford DL, et al.. 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated ? *Syst Biol* 50 :723–729.

Swofford D and Maddison W. 1987. Reconstructing Ancestral Character States Under Wagner Parsimony. *Math Biosci* 87 :199–229.

Szöllősi GJ, Boussau B, Abby SS, Tannier E, and Daubin V. 2012. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc. Natl. Acad. Sci. U.S.A.* 109 :17513–17518.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S. 2011. MEGA5 : molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28 :2731–2739.

Tateno Y, Takezaki N, Nei M, et al.. 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rates varies with site. *Mol Biol Evol* 11 :447–459.

Tekaia F and Yeramian E. 2006. Evolution of proteomes : fundamental signatures and global trends in amino acid compositions. *BMC Genomics* 7 :307.

Thorne JL, Kishino H, and Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol* 15 :1647–1657.

Tuffley C and Steel M. 1998. Modeling the covarion hypothesis of nucleotide substitution. *Math Biosci* 147 :63–91.

Whelan S and Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* 18 :691–699.

Williams PD, Pollock DD, Blackburne BP, and Goldstein RA. 2006. Assessing the accuracy of ancestral protein reconstruction methods. *Plos Comput Biol* 2 :e69.

Woese CR and Fox GE. 1977. Phylogenetic structure of the prokaryotic domain : the primary kingdoms. *Proc Natl Acad Sci U S A* 74 :5088–5090.

Xia X. 2013. DaMBE5 : a comprehensive software package for data analysis in molecular biology and evolution. *Mol Biol Evol* 30 :1720–1728.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites : approximate methods. *J Mol Evol* 39 :306–314.

Yang Z. 1995. A space-time process model for the evolution of DNA sequences. *Genetics* 139 :993–1005.

Yang Z. 1996a. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol* 11 :367–372.

Yang Z. 1996b. Maximum-Likelihood models for combined analyses of multiple sequence data. *J Mol Evol* 42 :587–596.

Yang Z. 2000. Maximum Likelihood Estimation on Large Phylogenies and Analysis of Adaptive Evolution in Human Influenza Virus A. *J Mol Evol* 51 :423–432.

Yang Z. 2006. *Computational Molecular Evolution*. Oxford University Press oxford university press inc., new york edition.

Yang Z. 2007. PAML 4 : Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* 24 :1586–1591.

Yang Z, Kumar S, and Nei M. 1995. A New Method of Inference of Ancestral Nucleotide and Amino Acid Sequences. *Genetics* 141 :1641–1650.

Yang Z and Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol* 19 :908–917.

Yang Z and Nielsen R. 2008. Mutation-Selection Models of Codon Substitution and Their Use to Estimate Selective Strengths on Codon Usage. *Mol Biol Evol* 25 :568–579.

Yang Z, Nielsen R, Goldman N, and Pedersen AMK. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155 :431–449.

Yang Z and Rannala B. 1997. Bayesian phylogenetic inference using DNA sequences : a Markov chain Monte Carlo method. *Mol Biol Evol* 14 :717–724.

Yang Z and Rannala B. 2006. Bayesian Estimation of Species Divergence Times Under a Molecular Clock Using Multiple Fossil Calibrations with Soft Bounds. *Mol Biol Evol* 23 :212–226.

Yang Z and Roberts D. 1995. On the Use of Nucleic Acid Sequences to Infer Early Branchings in the Tree of Life. *Mol Biol Evol* 12 :451–458.

Yap VB and Speed T. 2005. Rooting a phylogenetic tree with nonreversible substitution models. *BMC Evol Biol* 5 :2.

Zeldovich KB, Berezovsky IN, and Shakhnovich EI. 2007. Protein and DNA Sequence Determinants of Thermophilic Adaptation. *Plos Comput Biol* 3 :e5.

Zhang J and Nei M. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J Mol Evol* 44 :S139–S146.

Zimmer C. 2013. Bringing Extinct Species Back to Life. *National Geographic* 233 : :33–36.

Zoller S and Schneider A. 2012. Improving phylogenetic inference with a semiempirical amino acid substitution model. *Mol Biol Evol* 30 :469–479.

Zuckerkandl E and Pauling L. 1965. Molecules as documents of evolutionary history. *J Theor Biol* 8 :357–366.

Zuckerkandl E and Pauling LB. 1962. Molecular disease, evolution, and genetic heterogeneity. In *Horizons in biochemistry* pages 189–225. Academic Press New York m. kasha, b. pullman (eds) edition.